Lineshape fitting of iodine spectra near 532 nm

Tanaporn Na Narong (tn282@stanford.edu)

***Abstract*—Saturation absorption spectra of iodine were fitted to different lineshape functions. It was found that Voigt model gave the best fit to our data. Observed signal width w, Lorentzian width γ, and Gaussian width σ were extracted from all fitted spectra taken at different temperatures. We observed an increase in signal width and γ with temperature as expected. Next, for data taken at 21-25 °C from a depleted and an undepleted iodine cells, we applied Gaussian Discriminant Analysis, logistic regression, and SVM to classify the datasets taken from the two different cells. Using γ and σ as features, GDA and logistic regression were able to classify 83.58% of the data. However, SVM gave a better accuracy of 84.08% when using 2 principal components as features.**

I. Introduction

Absorption spectra of molecular iodine enable the study of its discrete vibrational energy bands over microwave and optical laser wavelengths. These line profiles, however, are often subject to various broadening mechanisms, which limit spectral resolution and precision of the measurements. One solution to this problem is to perform saturated absorption spectroscopy [1] to eliminate broadening due to Doppler effects and Gaussian velocity distribution of the molecules.

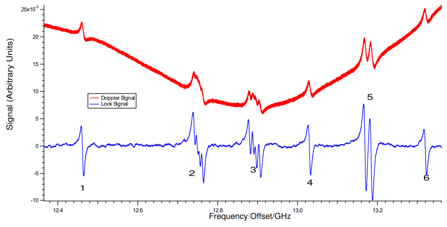

Fig. 1. (Red) Doppler-free spectra of iodine near 739nm. Multiple hyperfine lines are resolved within a single Doppler broadened structure. (Blue) First derivative of the fluorescence signal generated by lock-in electronics [2]

Each Doppler-free structure appears as a narrow absorption peak or dip on a baseline signal. Very often odd derivatives of the spectra are also taken to improve peak detection and flatten the baseline(see Figure 1). One common practice that was also used in obtaining datasets for this project is to extract third derivative of iodine spectra via phase sensitive detection to improve resolution. Currently our lab is interested in applications of iodine transition lines for time and frequency reference. For example, a laser frequency can be tuned and locked to one of the spectral lines with narrow linewidth. In order to build a highly stable experimental setup, detailed study of spectral lineshape and how these spectra can be affected by temperature and other experimental conditions should be pursued.

This project consists of 3 main stages. The first stage is to identify the lineshape of our spectra whether they could be best fit to Lorentzian, Gaussian, or Voigt function [2]. Using least square regression, we modeled the lineshape for all 740 spectra taken in the lab at different temperatures using two different iodine cells. We then extracted characteristic linewidths, namely observed signal width w, Lorentzian width γ, and Gaussian width σ, from each fit and plot them against temperature. Unweighted and weighted regression algorithms were used in this second stage to describe their temperature dependence. And finally, we specifically looked at data taken at 21-25 °C and applied classification algorithms namely GDA, logistic regression, and SVM to distinguish data taken from the 2 different cells. γ and σ widths were first used as features in this stage. We then also applied PCA to all 3 characteristic widths and extracted 2 principal components as the second set of features for this classification stage.

II. Related Work

A. Spectral lineshape functions

The 3 most fundamental lineshape functions used in this project are Lorentzian, Gaussian, and Voigt functions. The mathematical proofs and lineshape simulations were outlined in [2]. A summary is given below.

1) Lorentzian function: Assuming molecules are stationary, absorption spectra are expected to retain Lorentzian lineshape from lifetime broadening. The characteristic Lorentzian width γ denotes the FWHM of the line, which is proportional to the decay rate associated with each transition. However, actual linewidth observed is often larger due to pressure and power broadening. These broadening processes are temperature-dependent. The Lorentzian lineshape function centered at frequency ν_0 is given by

$$L(\nu) = I_0 \frac{\gamma}{(\nu-\nu_0)^2 + (\gamma/2)^2} \tag{1}$$

2) Gaussian function: When molecules are not stationary in the iodine cell, their Gaussian velocity distribution and Doppler effects when molecules absorb and re-emit light result in Doppler broadening. When this process dominates, the lineshape can be approximated as a Gaussian function centered at frequency ν_0 with Gaussian width σ.

$$G(\nu) = \frac{I_0}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\nu-\nu_0)^2}{2\sigma^2}\right) \tag{2}$$

In saturated absorption spectra as in Figure 1, excessive broadening from Doppler effects spans over many transition lines such that the lines would appear as a series of peaks on top of a Gaussian baseline. Directly from the fluorescence data, an algorithm for simultaneous baseline subtraction and spectral fitting was demonstrated in [3]. Their algorithm assumed the saturated absorption peaks were purely Lorentzian with relatively high intensities compared to the baseline. Unfortunately this did not seem to be the case for our noisy fluorescence data.

3) Voigt function: When subject to several broadening mechanisms, resultant lineshape could be modeled as a convolution of Lorentzian function (lifetime, pressure, and power broadening), and Gaussian function (Doppler broadening). There is no analytic functional form for Voigt lineshape but it can be calculated numerically from taking real part of Faddeewa function $w(z)$.

$$V(\nu) = L(\nu;\gamma) * G(\nu;\sigma) = \Re w(z) \tag{3}$$

where

$$w(z) = w(\frac{\nu + i\gamma}{\sigma}) = e^{-z^2}(1 + \frac{2i}{\sqrt{\pi}}\int_0^z e^{t^2}dt) \tag{4}$$

A MATLAB function to evaluate this Faddeewa function was written by Abrarov S.(2016) and made available to public [4]. Ruzi M.(2016) then wrote a MATLAB script for fitting several spectral lines simultaneously to Voigt lineshape function [5] using Abrarov's code. In this project we also made use of Abrarov's version of Faddeewa function in Voigt lineshape fitting but did not implement Ruzi's code directly as we ended up fitting the third derivative spectra instead of the original spectra.

Instead of working in frequency domain, an alternative method to perform lineshape fitting was proposed in [6]. Saarinen P.E. et al (1995) fitted data in time domain to the Fourier transform of the 3 lineshape functions above.

B. Phase sensitive detection: third derivative spectra

Using an experimental setup similar to [7] to obtain third derivative spectra, our laser frequency was modulated at 2kHz. This laser modulation resulted in modulated detected spectra, which was then demodulated at 3 times the modulation frequency, 6kHz. It can be shown mathematically that the Taylor expansion of the demodulated signal consists of the dominant term proportional to the third derivative of the absorption spectra plus small higher order terms.

Hongquan Li, another graduate student in our lab wrote a MATLAB script that calculated the third derivative spectra directly from third harmonic Fourier component. His script takes into account modulation (and demodulation) amplitude and frequency, and phase shift between laser modulation and reference signal. His code was used in lineshape fitting stage to demonstrate the difference between plainly taking third derivative of a spectrum and calculating exact solution, with the latter showing effects of higher order terms.

III. Dataset

Iodine saturated absorption spectra were measured by Hongquan Li and Prof Leo Hollberg, my research advisor, in Hollberg Lab, Department of Physics, Stanford University. Two sets of experiment were done using two different iodine vapor cells, one of them depleted and the other undepleted of iodine. There were in total 740 data files containing 396 spectra from the depleted cell at temperatures between 20-90 °C, and the other 344 spectra from the undepleted cell at temperatures between 20-45 °C.

A. Data preprocessing

Each data file corresponds to a single spectrum taken at one temperature during one cycle of laser frequency scan. There were 28035 entries in each file formatted into a 4005x7 matrix, whose columns correspond to 7 different signals measured simultaneously including PZT voltage controlling laser frequency scan, fluorescence signal, third derivative of the fluorescence, and temperature reading from a thermocouple.

We then cropped this matrix to keep only the ascending half of the laser scan cycle. A mid-point between the maximum and minimum of the third derivative signal was then centered at zero frequency. The fluorescence and third derivative signals from one of the data files are shown below.

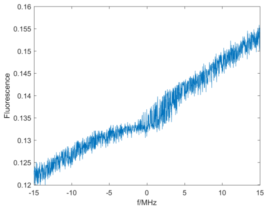

Fig. 2. Fluorescence signal on the centered frequency axis

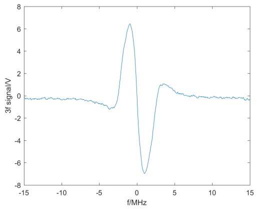

Fig. 3. 3rd derivative signal on the centered frequency axis

IV. Methods

A. Lineshape fitting

1) Locally Weighted Linear Regression: Locally weighted regression with Gaussian weight was applied to smoothen the spectra. Locally at query point ν we find θ that minimizes the weighted least square error with local weight $w^{(i)}$.

$$w^{(i)} = \exp\left(-\frac{(\nu - \nu^{(i)})^2}{2\tau^2}\right) \tag{5}$$

$$\theta = \underset{\theta}{\text{argmin}} \sum_{i=1}^{m} w^{(i)}(y^{(i)} - \theta^T\nu)^2 \tag{6}$$

At each query point, we perform weighted linear regression prioritizing neighboring data points. A straight line drawn locally would represent average trend of data points in the proximity quantified by the bandwidth parameter τ.

2) Ordinary Regression: For each lineshape function $f(\nu; b)$ described in section II, we determined fitting parameters $b = \{b_1, b_2, ...\}$ that minimize the least square error between data and calculated lineshape.

$$b = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{m} (y^{(i)} - f(\nu^{(i)}; b))^2 \qquad (7)$$

There are 4 fitting parameters for Lorentzian and Gaussian functions: characteristic width, signal strength, central frequency offset, and a constant vertical offset. For Voigt function there are 2 characteristic widths so there are 5 fitting parameters in total.

B. Temperature dependence of Lorentzian width

1) Locally Weighted Regression: We found that peak-to-peak height, $P2P$, of the spectra decreased dramatically with temperature. Assuming constant noise level, variance would be proportional to $1/P2P^2$. Hence local weight was chosen to be $P2P^2$, the square of peak-to-peak height at each temperature.

C. Vapor cell depletion comparison

For spectra taken at temperatures of 21-25 °C, we applied the following classification algorithms to distinguish data taken from two different iodine cells. One cell was depleted of iodine, and the other wasn't. The classification algorithms were applied to 2 different sets of features. A naive choice of features was Lorentzian width γ and Gaussian width σ. The other set of features were the two principal components of observed signal width w, γ, and σ.

1) Principal Component Analysis: PCA was used to reduce number of features from 3 to 2 by projecting the 3-dimensional feature vectors along 2 principal components.

Starting from mx3 matrix containing the 3 characteristic widths, we subtracted each feature column by its mean and then divide by its standard deviation to obtain a feature matrix with zero mean and variance of 1. We then calculated a 3x3 covariance matrix, its eigenvalues, and corresponding eigenvectors. The two principal components are the two eigenvectors with the two largest eigenvalues.

2) k-Nearest-Neighbor classification: Nonparametric kNN predicts an output locally by finding a weighted average output of the neighboring points. In this project, we assigned local weight $w^{(i)} = 1/d$ where d is the euclidean distance from the query point to its neighboring points.

3) Logistic Regression: We look for parameter θ that maximizes log-likelihood. Gradient ascent was used for parameter update.

$$l(\theta) = \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log (1 - h(x^{(i)})) \qquad (8)$$

where h(x) is a Sigmoid function

$$h(x) = \frac{1}{1 + \exp(-\theta^T x)} \qquad (9)$$

4) Gaussian Discriminant Analysis: Assuming the width features extracted from experiments with both iodine cells are normally distributed with the same covariance matrix but with different means, log-likelihood $l(\phi, \mu_0, \mu_1, \Sigma) = \sum_{i=1}^{m} \log p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma)$ is maximized when

$$\phi = \frac{1}{m} \sum_{i=1}^{m} 1 \left\{ y^{(i)} = 1 \right\} \qquad (10)$$

$$\mu_0 = \frac{\sum_{i=1}^{m} \left\{ y^{(i)} = 0 \right\} x^{(i)}}{\sum_{i=1}^{m} 1 \left\{ y^{(i)} = 0 \right\}} \qquad (11)$$

$$\mu_1 = \frac{\sum_{i=1}^{m} \left\{ y^{(i)} = 1 \right\} x^{(i)}}{\sum_{i=1}^{m} 1 \left\{ y^{(i)} = 1 \right\}} \qquad (12)$$

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \qquad (13)$$

These parameters are then used to calculate the Gaussian contours $p(x|y)$. $p(y|x)$ is evaluated using Naive-Bayes assumption, and the decision boundary is where $p(y = 1|x) = p(y = 0|x) = 0.5$.

5) Support Vector Machines: Linear SVM model separates data with labels $y = 1$ or $y = -1$. In 2 dimensions, margin between the lines $w \cdot x - b = 1$ and $w \cdot x - b = -1$ is maximized, which is equivalent to minimizing $||w||$ subject to $y^{(i)}(w \cdot x^{(i)} - b) \geq 1$ for all i.

V. Results & Discussion

A. Lineshape fitting

1) Fluorescence signals: After smoothing the data using bandwidth $\tau = 1$, Linear baseline and Lorentzian curve fitting to the residual dip were performed on fluorescence data. However, modulation present in the fluorescence data resulted in 80 times higher error-to-signal ratio than fitting the 3rd derivative of Lorentzian lineshape to third derivative spectra directly.

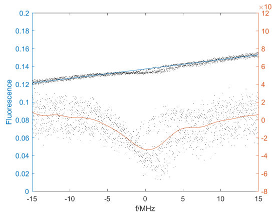

Fig. 4. (Blue) Fluorescence signal with the linear baseline fit. (Red) Residual signal showing the absorption dip

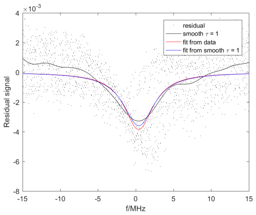

Fig. 5. Lorentzian fit (Red) directly on residual signal, and (Blue) smooth signal using τ = 1

TABLE I
Fitting errors from Lorentzian model

Spectra	RMS Error	I_0	Error/I_0
Fluorescence	0.0016	0.0048	0.3333
3rd derivative	0.2280	54.57	0.0041

2) Third derivative signals: We decided to use third derivative spectra instead of fluorescence data for lineshape fitting for better signal-to-noise ratio. We first performed Gaussian weight smoothing. After experimenting with different values of τ, we found that $\tau = 0.2$ was an optimal choice that helped reduce fitting error without overly reducing signal height and width compared to the shape of original signal.

On smooth spectra, we used least square regression to fit the lineshape using 4 different models. 3 of them were the 3rd derivative of Lorentzian, Gaussian, and Voigt functions, and the last one was calculated directly from demodulating a Lorentzian signal at 3 times the modulation frequency.

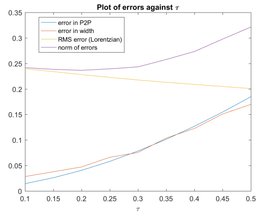

Fig. 6. Fitting errors when using different smoothing parameters τ

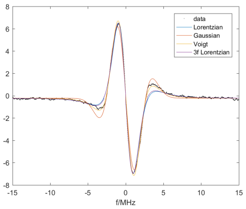

Fig. 7. Lineshape functions comparison on a 3rd derivative spectrum

TABLE II
FITTING ERRORS FROM DIFFERENT LINESHAPE MODELS

Model	RMS Error	%Error in P2P	%Error in w
Lorentzian	0.228	4.7	3.6
Gaussian	0.248	0.2	12.6
Voigt	0.126	3.5	6.3
Lorentzian(exact)	0.232	4.6	0.9

The metrics used in selecting both τ and the best lineshape model were RMS error, percentage differences between fitted $P2P$ and signal width w to their corresponding values extracted from raw data. It turned out Voigt model gave the least RMS error whereas Gaussian and Lorentzian functions best preserved signal height $P2P$ and width w respectively. The errors shown in the table above were averaged over random samples of 20 spectra.

Because all spectra were taken in the same experiment under the same conditions, we assumed Voigt lineshape model would produce the least errors when fitting the remaining data. The histograms in figures 8 and 9 show separate error distributions from Voigt fit on data taken from the depleted cell and undepleted cell. From each Voigt fit we extracted w,γ, and σ for later stages of this project.

B. Temperature dependence

As temperature increased, overall signal got smaller for both depleted and undepleted cell. The signal completely disappeared above 45°C for the undepleted cell. In figure 11, Lorentzian width γ increased with temperature. Gaussian width σ on the other hand, seemed to decrease with temperature in the depleted cell but remained roughly constant in the undepleted cell.

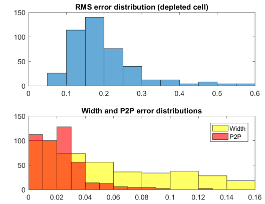

Fig. 8. Error distributions for depleted cell data

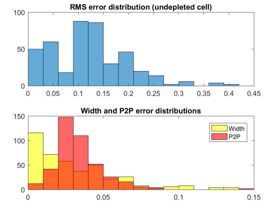

Fig. 9. Error distributions for undepleted cell data

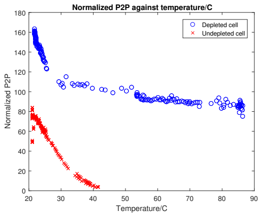

Fig. 10. Plot of signal peak-to-peak against temperature(°C)

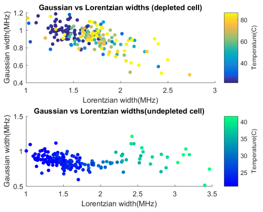

Fig. 11. Scatter plot of σ and γ shown with temperature color bar

We then applied weighted and unweighted linear regression to the plots of Lorentzian width against temperature. Surprisingly the slope in the undepleted cell plot was up to 10 times higher than that for the depleted cell. Negative intercepts may have suggested that this linear dependence no longer applied at lower temperatures as we expected γ to remain positive at all temperatures.

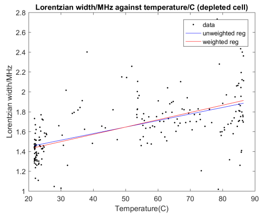

Fig. 12. Plot of γ/MHz against temperature(°C) for depleted cell

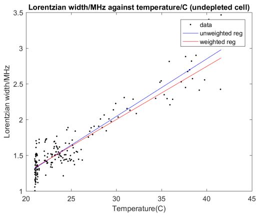

Fig. 13. Plot of γ/MHz against temperature(°C) for undepleted cell

TABLE III
WEIGHTED AND UNWEIGHTED REGRESSION RESULTS

Cell	Weighted	Intercept(MHz)	Slope(MHz/°C)	RMS error
Depleted	No	1.3145	0.0066	0.2359
Depleted	Yes	1.2767	0.0074	0.2367
Undepleted	No	-0.3716	0.0805	0.1654
Undepleted	Yes	-0.2310	0.0744	0.1696

C. Vapor cell classification

Classification algorithms were applied to datasets taken from the depleted and undepleted cell at temperatures between 21-25°C. We specifically picked this range of temperature in which the characteristic widths w,γ, and σ were roughly normally distributed, and temperature dependence over this small range could be neglected. We carried out this classification stage using 2 different sets of features.

1) Lorentzian width and Gaussian width as features: k-Nearest Neighbor using $1/distance$ as weight gave a baseline learning accuracy of 81.09% when applied to the whole dataset of 402 samples. Assuming normally distributed data, GDA gave a slightly higher accuracy of 83.58% which was the same as achieved by logistic regression.

We first implemented mini-batch gradient ascent to train the logistic regression model. Learning rate was chosen to be 1 after obtaining worse results from larger values. The batch sizes tried were 5, 10, 20, 40, and 80. It was found that training with smaller batch sizes ran and converged more slowly; batch size of 5 did not reach convergence even after 200,000 epochs. We ended up switching to normal gradient ascent that went through entire dataset for each parameter update. Same accuracy as GDA was achieved in around 100,000 iterations for each fold of 5-fold cross validations.

SVM model for linear classification was found using built-in MATLAB function `fitcsvm`. We did not expect excellent results from SVM because the two clusters weren't separable. Learning accuracy from SVM was 78.61% which was slightly lower than GDA and logistic regression accuracies.

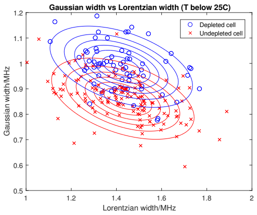

Fig. 14. Plot showing GDA Gaussian contours for depleted and undepleted cell data

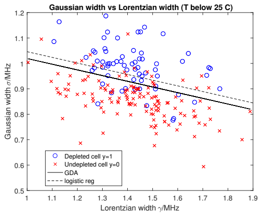

Fig. 15. Plot showing decision boundaries from GDA and logistic regression on γ and σ

TABLE IV
LEARNING ACCURACY(%) USING DIFFERENT ALGORITHMS

Model	5-fold Training	5-fold Test	Entire set
kNN	-	-	81.09
GDA	83.23	83.00	83.58
Logistic regression	83.60	83.00	83.58
SVM	77.58	77.75	78.61

2) Features extracted from PCA: 2 principal components were extracted from 3 features: signal width w, γ, and σ. The new features V_1 and V_2 were the corresponding eigenvectors of the covariance's matrix largest eigenvalues of 1.2535 and 1.7427 respectively.

$$V_1 = 0.9051w + 0.4252\gamma \tag{14}$$

$$V_2 = -0.6759\gamma + 0.7370\sigma \tag{15}$$

We then applied classification algorithms to the new features. kNN performed slightly worse than when trained on original features. GDA and logistic regression had the same learning accuracy of 83.58% as before. Surprisingly SVM model gave the highest learning accuracy of 84.08% when applied to the entire data set When data are projected along the principal components, variance is maximized such that there could be less overlap between the two clusters. Therefore, SVM performed better with PCA features.

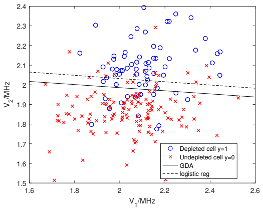

Fig. 16. Plot showing decision boundaries from GDA and logistic regression on PCA features V_1 and V_2

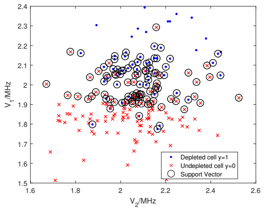

Fig. 17. Plot showing support vectors from SVM model

TABLE V
LEARNING ACCURACY(%) USING DIFFERENT ALGORITHMS

Model	5-fold Training	5-fold Test	Entire set
kNN	-	-	79.10
GDA	83.73	83.00	83.58
Logistic regression	83.48	83.50	83.58
SVM	82.80	80.75	84.08

VI. Conclusion

Iodine spectral lineshape could be described by Voigt function with characteristic Lorentzian width γ and Gaussian width σ. For temperatures between 20-90°C, Lorentzian width increased roughly linearly with temperature. For spectra taken at 21-25°C from two different iodine cells, machine learning algorithms were able to classify the data at about 84% accuracy using γ and σ as features or the principal components as features.

VII. Future Work

Detailed error analysis on the lineshape model should be carried out so that model errors can be compared and correlated to estimated experimental errors. Hopefully with lower noise it would be possible to simultaneously fit the fluorescence data and its third derivative to verify the lineshape model and fitting parameters. Third derivative of Voigt model as the lineshape function may be replaced by exact calculation of third harmonic Fourier component, similar to what we had demonstrated with Lorentzian lineshape. Another modification to lineshape fitting would be to include the phase ϕ in phase-sensitive detection as another fitting parameter as it might not have been zero as we had assumed.

Studies of temperature dependence and cell depletion comparison could be improved significantly with better lineshape modeling, more data points, and less experimental noise.

References

[1] Magyar, J. A., & Brown, N. (1980). High resolution saturated absorption spectra of iodine molecules 129I2, 129I127I, and 127I2 at 633 nm. *Metrologia*, 16(2), 63.

[2] Chew, A. (2008). Doppler-free spectroscopy of iodine at 739nm. *undergraduate thesis*, University of Maryland.

[3] Han, Q., Xie, Q., Peng, S., & Guo, B. (2017). Simultaneous spectrum fitting and baseline correction using sparse representation. *Analyst.*

[4] Abrarov, S. (2016, July 10). The Voigt/complex error function (second version). Retrieved November 30, 2017, from https://www.mathworks.com/matlabcentral/fileexchange/47801-the-voigt-complex-error-function–second-version-

[5] Ruzi, M. (2016, June 27). Voigt line shape fit. Retrieved November 30, 2017, from https://www.mathworks.com/matlabcentral/fileexchange/57603-voigt-line-shape-fit

[6] Saarinen, P. E., Kauppinen, J. K., & Partanen, J. O. (1995). New method for spectral line shape fitting and critique on the Voigt line shape model. *Applied spectroscopy*, 49(10), 1438-1453.

[7] Kobayashi, T., Akamatsu, D., Hosaka, K., Inaba, H., Okubo, S., Tanabe, T., ... & Hong, F. L. (2015). Compact iodine-stabilized laser operating at 531 nm with stability at the 10 12 level and using a coin-sized laser module. *Optics express*, 23(16), 20749-20759.

[8] Parsons, R. W., Metchnik, V. I., & Dyne, R. J. (1968). The collision broadening of spectral lines. *Australian Journal of Physics*, vol. 21, p.13