# ”The smell of fear” - In Search of Patterns Between Emotions and Breath Composition - Project Proposal

Jason Tam[1], Musa Abdel-Rahman[2], Catherine Liu[3]
Supervised by Jörg Wicker

***Abstract*— As Volatile Organic Compounds (VOCs) components of gas excreted from plants have shown to play a role in their communications, the exploration to determine the existence of equivalent phenomenon between humans have drawn research interests. Gas data have been captured in movie theatres for the purpose of determining potential associations and relationships between human emotion and breath composition. This paper proposes unsupervised methods and their applications to the collected data, for determining potential associations between gas data and scene types that were presented to the audience by the movie.**

## I. INTRODUCTION

An important process of many biological organisms is gas excretion. The gas excretions of humans compose of Carbon Dioxide and various Volatile Organic Compounds (VOCs) such as isoprene and acetone. These VOCs have been found to play an important role in the communication between plants, and even between plants and animals, however we do not know if these VOCs have any impact on human communication. Research within the field to understand how these VOCs interact or are correlated with human emotions have been attempted by [1].

The intention of this study is to continue on from the accomplishments mentioned in [1]. The current plan is to further explore the potential patterns/associations that are intrinsically present in the collected dataset, such as determining if there are molecular masses with high detected counts in multiple movie scene labels or to determine if there are association rules between the movie scene labels and ranges of molecular masses (instead of the individual channel per molecular mass size in the data). This can be achieved by starting with general pattern mining strategies with unsupervised learning techniques, and may apply supervised learning methods for finer details of the resulting patterns if feasible. This paper explores techniques that are potentially useful for various analysis possibilities of the data, such as using *Dynamic time warping* to align different sessions of the same movie, or clustering techniques that may reveal high volume detection in certain molecular mass channels are trigged by certain scene types.

## II. DATA STRUCTURE AND NORMALIZATION

The data is obtained from the Kaggle repository named ”The smell of fear” [2], which is the same dataset from [1].

[1]jtam013@aucklanduni.ac.nz
[2]mabd292@aucklanduni.ac.nz
[3]cliu129@aucklanduni.ac.nz

The dataset contains gas data collected from a cinema theatre in Mainz, Germany over the time period from 2013-12-18 to 2014-01-16. The detector contains 431 gas channels, each representing a particular molecular mass value as well as a dedicated channel for $CO_2$. It records a measurement into each of these channels every 30 seconds. The recording was continuous over the entire time period, hence it contains measurements from movie sessions as well as times when the theatre is empty. Other associated data are also present in the dataset, such as screening times and attendance of the movies played in the theatre over this time period, and classified scene types (such as ’Action’, ’Comedy’) per 30 second interval for six movies. These movies were chosen for their similarities within theme and production (such as ’The Hunger Games’ and ’The Hobbit’), or for their ability to evoke drastic reactions from the audience during certain scenes that may not have otherwise been stimulated from a more calm movie (such as ’Machete Kills’ and ’Paranormal Activity’).

The quantity detected in each of the gas channels is assumed to be correlated to the audience size in the theatre. A normalization procedure is performed to assist aligning gas data from different sessions of the same movie. Data from the $CO_2$ channel is used to determine time periods appropriate for determining background levels. Data collected in the early hours from 0400 to 0900 between the dates of 2013-12-30 to 2014-01-07 appear to be the quietest, and the average of these values seem to be a good representation of background counts. A comparison of the data collected during these times to others is displayed in Figure 1.

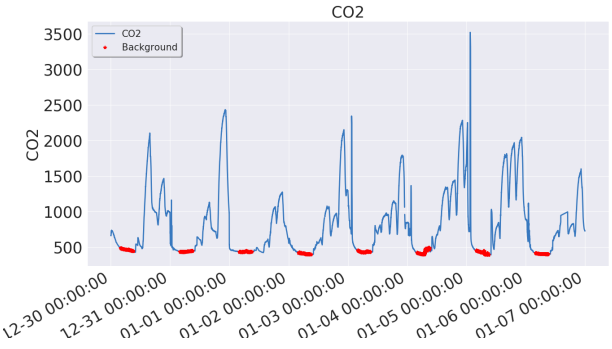


Fig. 1. Carbon Dioxide background of ’Hobbit 2’

As the relationship between $CO_2$ and human breath is well understood, these time periods can be assumed to also

be appropriate for determining background counts in other channels. The normalization process involves subtracting the averaged background counts from each of the session and dividing by the respective attendance. The resulting value at each of the 30 second interval is representative of the average contribution per person in the audience for that gas channel in that specific movie session. A comparison of the normalized $CO_2$ channel for all sessions of the 'The Hunger Games - Catching Fire' movie is shown in Figure 2.

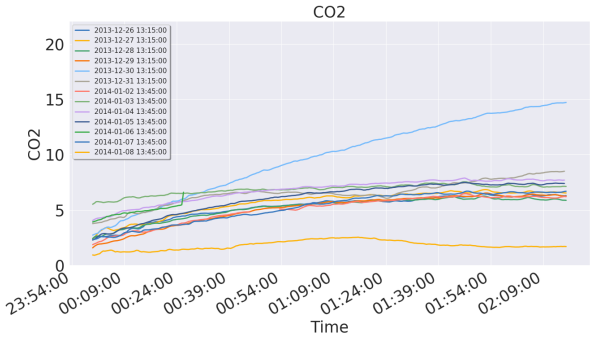


Fig. 2. Statistics of the Carbon Dioxide channel from all sessions of 'The Hunger Games - Catching Fire'.

For each of the movies; morning, afternoon, and evening sessions were screened, every day of the week. The attendance numbers for these sessions varied drastically from 2 through to 227. This meant that plotting the raw data showed large variances in channel values at each time stamp as depicted in the graph below. However, it can be seen that a general trend was still followed regardless of the number of people in the cinema, as displayed in Figure 3.

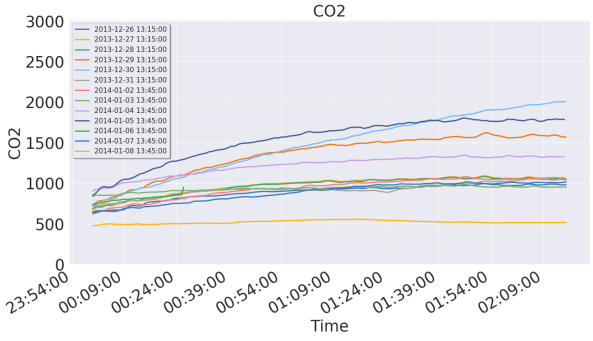


Fig. 3. Raw data of the Carbon Dioxide channel from all sessions of 'The Hunger Games - Catching Fire'.

Dynamic Time Warping (DTW) will be implemented after the normalization process to ensure the alignment of the multiple sessions within each movie. Measurements for gas concentrations were taken every 30 seconds, however there may be the possibility that through human or equipment error, misalignment of the data may occur. This could be through a delayed response in human reaction, or the data sensors that pick up measurements may have a slight lag.

It is assumed that within the enclosed cinema, that the human breath is the most dominant component within the readings from the mass spectrometer and would be drastically larger than any gases that may be emitted through any external factors such as beverages or drinks movie goers are ingesting (e.g., popcorn and alcohol) [3]. It was found that the best way to work with the data was to have an overall average value that was allocated for every channel at every time frame, for each movie. This was done by adding together the normalized values of a channel at a particular time stamp within all the session of the movies and averaging this across the number of movie sessions. It was important to manipulate the data so that it showed the average contribution per person for that gas channel as this mitigated any discrepancies that may occur in attendance numbers between the different movies.

Following on from this, a new data frame was created that had the ratio of each particular gas channel against the total sum of all the gases at each time frame. This was done so that it would be computationally easy to understand and analyze when a certain molecular channel was 'excited', (i.e., an increase in volume value) as this could be compared through ratios and percentages, as opposed to the actual gas values. Having the actual values would be harder to interpret what would be considered a 'large' increase in volume when using the magnitude of the values.

Within the datasets that contained label information for each of the movies, a vector was assigned accordingly based on the labels that were active during each time frame. Once these were all assigned, the screening times of the movies were married up to the new data frame, and the appropriate movie and its 'labels' vector were added to this dataset. From here, the data could be used in a way that was independent of the time stamp. This was simply done by filtering the data to have only the rows that contained a 'labels' vector present. By doing this, the resulting dataset was created that contained a column of vectors that represented labels active when the information was collected, and corresponding columns that had the ratio of the molecular channel concentration present against the total molecular channels at those times.

## III. METHODS AND ALGORITHMS

Clustering data into groups based on similarity is often one of the most important steps in any data-analysis application. The goal of data clustering is to identify homogeneous groups or clusters from a set of objects. In other words, data clustering aims to divide a set of objects into groups or clusters such that objects in the same cluster are more similar to each other than to objects from other clusters [4]. Based on the data preparation that we are making ,we will implement HCA Clustering algorithm to assign the "labels" that have the same behavior, which would produce the same effect on the gases.

HCA techniques attempt to separate data into specific groups, based on a similarity measure. Initially each data point represents its own cluster, then the threshold for the decision when to declare two or more objects to be members

of the same cluster is lowered incrementally [5] [6]. This strategy will produce hierarchical representations in which the clusters at each level of the hierarchy are created by merging clusters at the next lower level. At the lowest level, each cluster contains a single observation. At the highest level there is only one cluster containing all of the data [4], and that can be graphically displayed in a *Dendrogram Diagram*. This diagram will represent a Binary tree that reflect the *agglomerative* and the *divisive* approaches, which are used to show the similarity and dissimilarity between groups [4] [5]. A visual illustration of this method is presented in Figure 4.

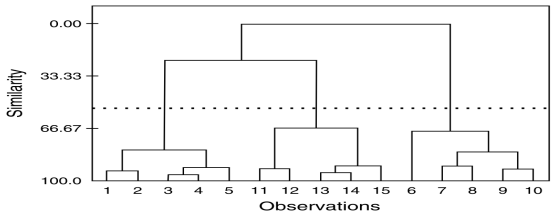


Fig. 4. Dendogram illustrating HCA data clustering [5]

One method of hierarchical clustering analysis that will be used will be a bottom up/agglomerative approach. This will look into the 'labels' vectors column and cluster based on the frequency of the label within the vector. Each vector within the column comprises of a 42 bit string, where '1' indicates the label was active during that time, with the position of the '1' indicating the label itself. Each label vector will start off as its own cluster, and from here it will find the best pair to merge into a new cluster. The distance matrix will comprise of each of the 'labels' vectors and distance can be computed by treating the vectors as an n-cube. In this case, it could be thought of as a 42-cube or $Q_{42}$ with bit strings that only differed by one position to being of distance 1. To compare cluster similarities, for example, Ward's linkage may be used as it is less susceptible to noise and outliers. Within the data, the assumption has been made that human breath is the largest component within the readings, however the noise generated by miscellaneous external environmental factors (such as various food or drinks) may still play a factor, so by utilizing Ward's linkage, this may mitigate any noise that may occur. From here, an appropriate number of clusters can be chosen, and the cluster each 'labels' vector belongs to is assigned and added to the data set. Using an aggregative method (such as using the average or sum of the values), we can obtain one value for each of the cluster groups within every molecular channel. These values will be statistically compared (both within the molecular channel itself and against other channels) to see if there are any similarities, differences or patterns within the data.

## IV. CONCLUSION

The main proposed research is to explore the potential patterns and associations that are intrinsically present in the collected dataset, through the use of unsupervised learning techniques.

To achieve these objectives, the data was manipulated using DTW and statistics methodologies that meant an average, normalized contribution value per person for each molecular channel at each time frame for every movie was obtained. A new data frame was then created that had ratios of each of the channels against the sum of the total channels at a particular time stamp, and the labels associated with that time frame were also added through the use of a 42 bit string. From here, the data set can be utilized independent of time.

Hierarchical cluster analysis of $n$ objects is an algorithm that merges objects that have close similarities. The similarities between clusters of objects can be defined in several ways of linkage techniques. Either rows or columns of a matrix can be clustered, in each case we can choose the appropriate similarity measure that we prefer. The results of a cluster analysis is a binary tree, or dendrogram, with $n$-1 nodes. The branches of this tree are cut at a level where there is a lot of 'space' to cut them, that is where the jump in levels of two consecutive nodes is large. HCA has several advantages, for example it is conceptually simple, does not require the number of clusters k in advance and it is less sensitive to noise in the data set.

## REFERENCES

[1] J. Williams, C. Stönner, J. Wicker, N. Krauter, B. Derstroff, E. Bourtsoukidis, T. Klpfel, and S. Kramer, "Cinema audiences reproducibly vary the chemical composition of air during films, by broadcasting scene specific emissions on breath," *Scientific Reports*, vol. 6, no. 25464, 2016.

[2] J. Wicker, "Kaggle repository - 'the smell of fear'," 2018. Accessed: 2018-09-17.

[3] J. Williams and J. Pleil, "Crowd-based breath analysis: assessing behavior, activity, exposures, and emotional response of people in groups," *Journal of Breath Research*, vol. 10, no. 3, p. 032001, 2016.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *Unsupervised Learning*, pp. 485–585. New York, NY: Springer New York, 2009.

[5] S. M. Scott, D. James, and Z. Ali, "Data analysis for electronic nose systems," *Microchimica Acta*, vol. 156, pp. 183–207, Dec 2006.

[6] K. Ramasubramanian and A. Singh, *Machine Learning Theory and Practices*, pp. 219–424. Berkeley, CA: Apress, 2017.