1 Introduction

olfactory receptor neurons transduce chemical stimuli into electrical signals to detect odorants. These signals are relayed to the olfactory bulb (OB) or the antennal lobe (AL). In the AL and OB, neural codes for odors take the form of spatiotemporal patterns, which contain various information about the odorants. These patterns can be measured using Electroantennagraphy (EAG), a technique for measuring the average output of an insect antenna to its brain for a given odor. Responses to various odorants, parts of the antenna, mixes of odorants and concentrations were measured and analyzed using machine learning. Here, an analytic method was proposed and examined to identify odorants and concentrations and to find the composition of mixes of odorants.

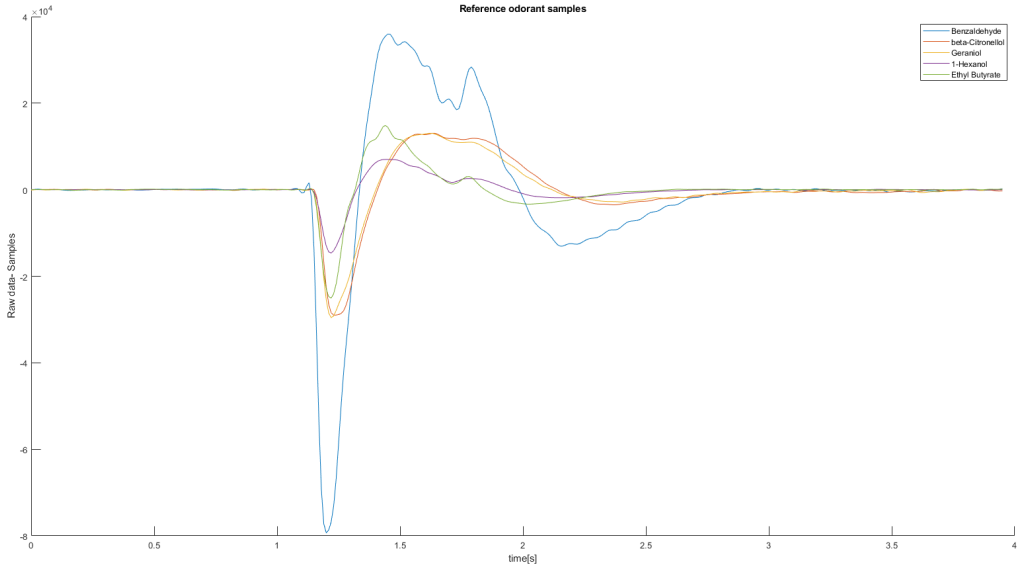

Figure 1: Reference sample of each odorant.

2 Methods

2.1 Correlation

To identify an odorant, a correlation between odorant responses is used:

$$Odorant_{correlation} = \frac{\left|\int \psi(t) \cdot \psi_{odorant}(t) dt\right|^2}{\int |\psi(t)|^2 \, dt \int |\psi_{odorant}(t)|^2 \, dt} \tag{1}$$

Here ψ is the analyzed odorant response and $\psi_{odorant}$ is the reference odorant response. The correlation with highest value is the identified odorant.

2.2 t-distributed stochastic neighbor embedding, t-SNE

t-SNE is an algorithm for dimensionality reduction that is well-suited to visualizing high-dimensional data. The idea is to embed high-dimensional points in low dimensions in a way that respects similarities between points. Nearby points in the high-dimensional space correspond to nearby embedded low-dimensional points, and distant points in high-dimensional space correspond to distant embedded low-dimensional points. (Generally, it is impossible to match distances exactly between high-dimensional and low-dimensional spaces.) The tsne function creates a set of low-dimensional points from high-dimensional data. Typically, you visualize the low-dimensional points to see natural clusters in the original high-dimensional data.

2.3 Principal Components Analysis, PCA

The main purpose of PCA is to find the subset of features of our dataset that best encaptures information on the whole data so that we can reduce dimensions with minimal loss of information. Several techniques exist for dimensionality reduction, PCA approaches this task by identifying principal components, that are linear combinations of the original features. These components are extracted so that the first principal component encaptures maximum variance in the dataset, the second encaptures the remaining variance while being uncorrelated to the first, and so on.

3 Results

3.1 Single Odorants

To begin the analytic analysis I used data of 5 odorants: Benzaldehyde,β-Citronellol,Geraniol,1-Hexanol and Ethyl Butyrate, at a similar concentration, where the AEG is taken from the entire antenna, as seen in fig.1. First, I tried to work with the raw data to understand the problem better. Taking one sample of each of the

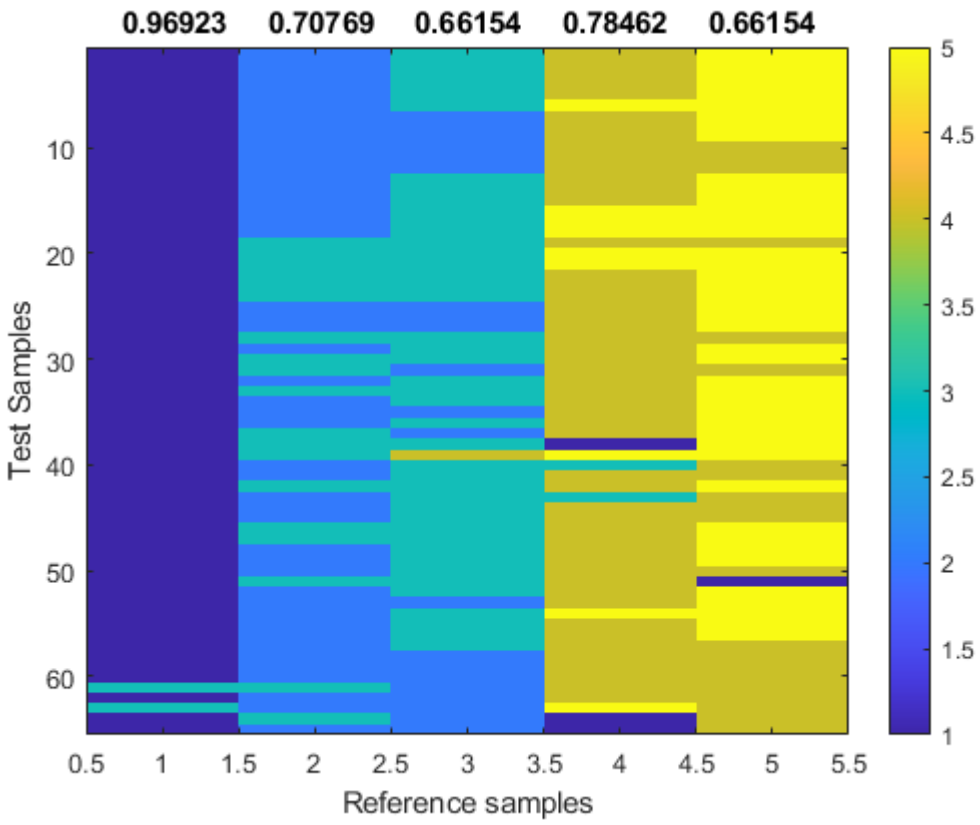

Figure 2: Correlation results for single odorants.The matrix is a 5X65 matrix organized so that column 1 includes all the samples of odorant 1, and so on. The color of each cell indicates the maximal correlation of the sample. The numbers in the title are the overall success rate for each of the odorants. Here, the colors indicate the indices of the odorants, so that: 1='Benz',2='Cit',3='Geran',4='Hex',5='Ethyl'.

5 odorants as a reference and correlating it to the rest of the samples using eq. 1 I get a different success rate for each odorant, where the lowest is 66% and the highest is 97%, as shown in fig. 2. I tried various methods to improve this result. The more trivial ones, such as: normalizing the data, filtering it with various filters to extract data from certain frequencies of interest, using different areas of the signals- positive or negative etc. Later, I tried to use spectral data. I used various transforms and tried to extract interesting features for the correlation, for example: Fast Fourier transform, Discrete Fourier transform with second-order Goertzel algorithm, Discrete cosine transform, Hilbert transform, Fast Walsh-Hadamard transform. I also tried time-frequency analysis: Empirical mode decomposition, Hilbert-Huang transform, Variational mode decomposition, Short-time Fourier transform using specific data from the result. I tried those methods to find similarities between different samples of the same odorant and differences between odorants. All those methods didn't improve the result.

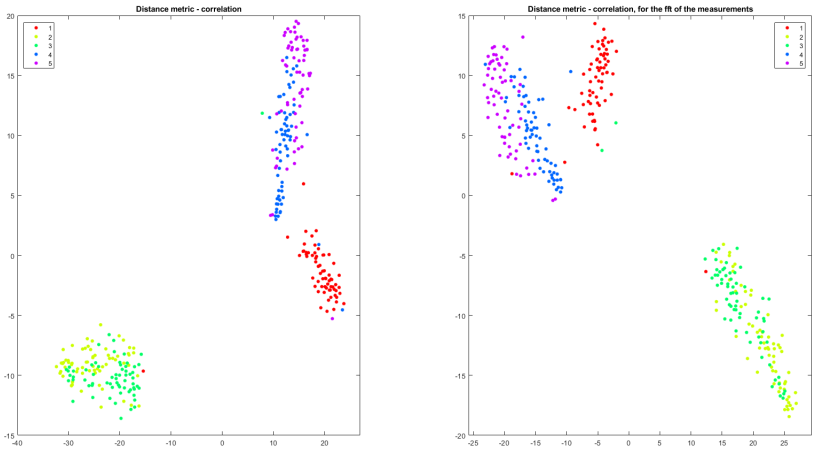

Figure 3: t-SNE. The distance metric is correlation. Left: For data normalized to values between 0-1. Right: For data normalized to values from 0 to 1, Fourier transformed and then taking only the real part.The indices of the odorants are: 1='Benz',2='Cit',3='Geran',4='Hex',5='Ethyl'.

3.2 Mixes of Odorants

Two mixes of odorants were recorded, consisting of 2 odorants each. Using the correlation method and choosing the top 2 maximal correlations for each of the mixes I didn't succeed in extracting the correct composition of those mixes. I also used t-SNE following various algorithms and distance metrics to try to show the difference between odorants. The goal was to separate them enough so that mixes of the odorants could be placed on a 2D plane in a way that clarifies what odorants they are a mix of. In case the distance was far enough between the odorants in a 2D plane, maybe that data could have been used to create EAGs of those odorants. The result was not distinct enough to do so as seen on fig.3. Last, I used PCA on the mixes and the reference odorants and calculated the distance and correlation between those vectors. For signals normalized to values between 0-1 the results were better. First I averaged over all the samples of each mix and odorant. For the mix of Benzaldehyde and β-Citronellol the shortest distances were those of Benzaldehyde and 1-Hexanol. For the mix of 1-Hexanol and Ethyl Butyrate the shortest distances were gladly those of 1-Hexanol and Ethyl Butyrate. Calculating for each of the samples of the odorants and mixes the shortest distances for the mix of Benzaldehyde and β-Citronellol were correct only 3%, the shortest distances for the mix of 1-Hexanol and Ethyl Butyrate were correct 68%.

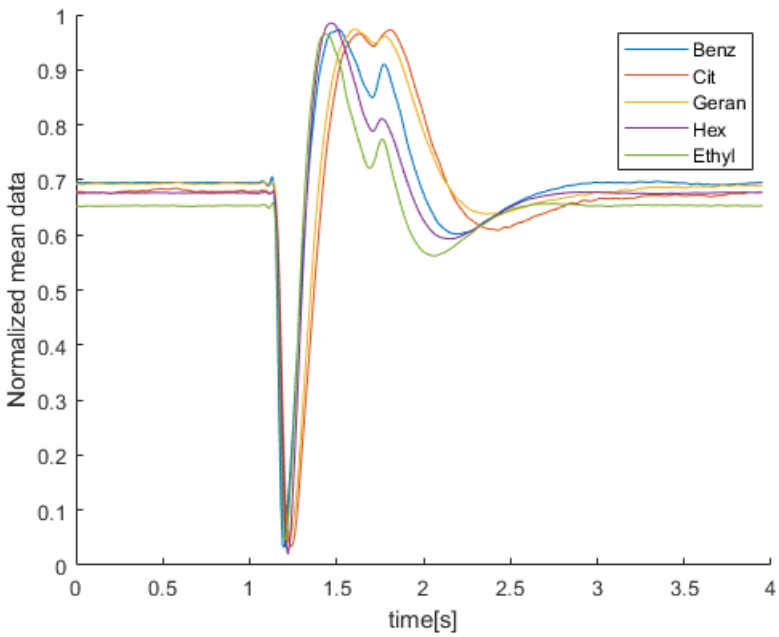

Figure 4: Reference average of samples of each odorant.

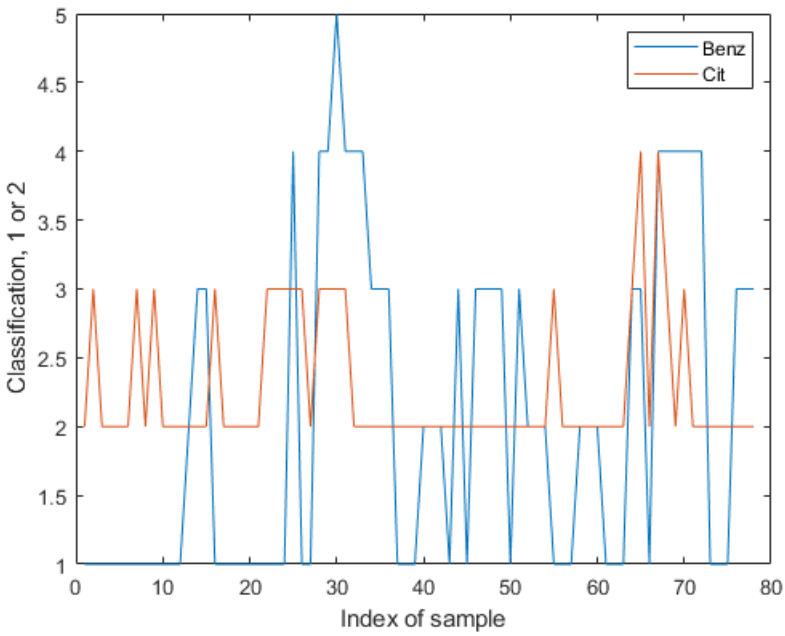

Figure 5: Classification of Benzaldehyde and β-Citronellol from both halves of the antenna. Correlation of the reference samples and the samples of the halves. The samples are normalized to values between 0-1. The indices of the odorants are: 1='Benz',2='Cit',3='Geran',4='Hex',5='Ethyl'.

3.3 Antenna Halves

As reference, instead of taking the first sample of each odorant I used the average of the normalized samples, as shown in fig. 4. Then, I used the normalized measurements of Benzaldehyde and β-Citronellol from both halves of the antenna as the samples for test and correlated each of them with the reference samples. The success rate for Benzaldehyde was 76% while the success rate for β-Citronellol was only 50%, as seen in fig. 5. Using various methods for preprocessing of the data didn't quite improve the results. The simplest one being

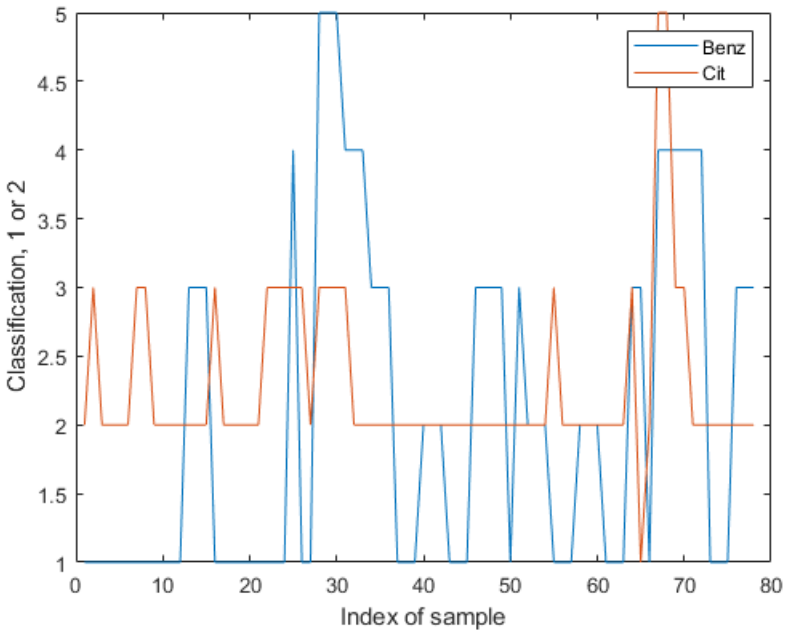

Figure 6: Classification of Benzaldehyde and β-Citronellol from both halves of the antenna. Correlation of the reference samples and the samples of the halves. The samples are normalized to values between 0-1, fourier transformed and then only the real part is taken. The indices of the odorants are: 1='Benz',2='Cit',3='Geran',4='Hex',5='Ethyl'.

data normalized to values from 0 to 1, Fourier transformed and then taking only the real part. The success rate for Benzaldehyde was 74% while the success rate for β-Citronellol was only 51%, as seen in fig. 6. Other methods as mentioned above didn't improve the results as well.

3.4 Odorant Concentrations

As reference I used the average of the samples of each odorant. First, I used data of 4 concentrations to check if correlation could identify which odorant it is, aiming to continue and find the concentration value if this step succeeds. I used 4 concentrations of Geraniol and 12 samples from each concentration. A very interesting result is that given 12 samples, in all concentrations only the last 3 (from the same subject) gave good results, identifying the odorant. As seen in fig. 7 the success rate is 31%. For single samples that aren't averaged the success rate is lower, 22%. For Benzaldehyde using 6 concentrations and 27 samples for each of them the corresponding success rates are 38% and 48%.

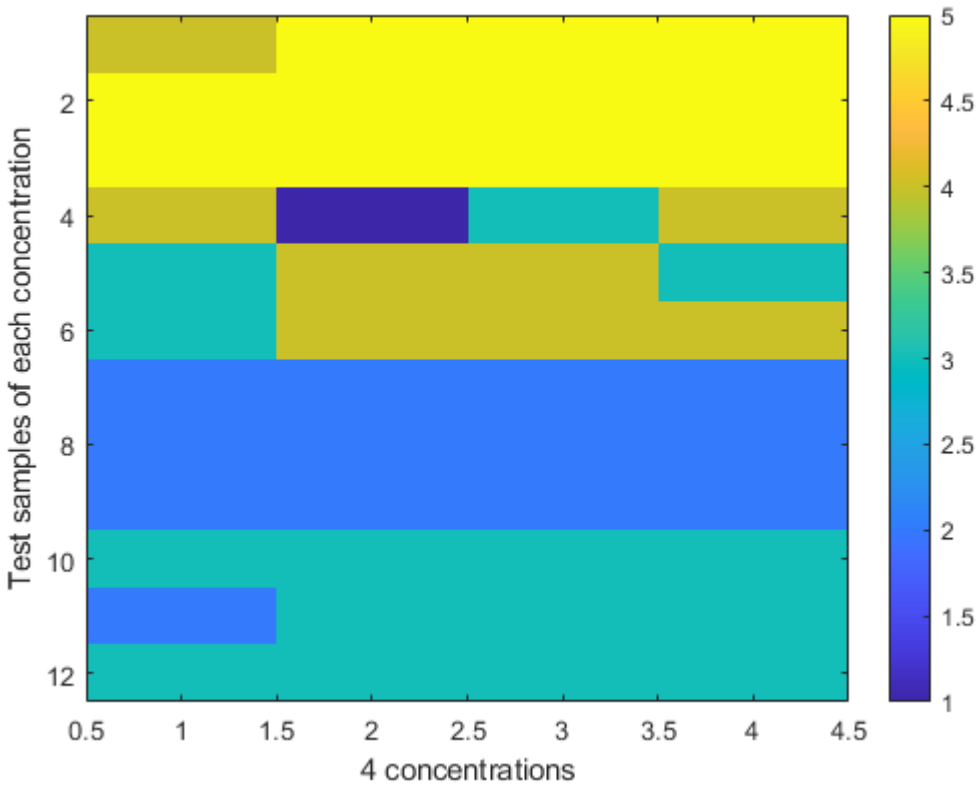

Figure 7: Correlation results for concentrations of Geraniol. We get a matrix of 4X12, 4 concentrations and 12 samples from each concentration. The color of each cell indicates the maximal correlation of the sample. The numbers in the title are the overall success rate for each of the odorants. Here, the colors indicate the indices of the odorants, so that: 1='Benz',2='Cit',3='Geran',4='Hex',5='Ethyl'.

4 Discussion

The analytic approach is less efficient here due to the variations between samples. Some of the odorants are more distinct than others and are easier to classify and identify even from mixes. Several pre-processing methods didn't improve the results of the methods used to analyze the odor data. An interesting observation is that odorants with indices 2,3 are similar to each other and also 3,4 are similar to each other. To conclude, machine learning is more fit for detection of odors and for their classification.

5 References

1. B, Joseph J, Tang J, Stopfer M. Temporally diverse firing patterns in olfactory receptor neurons underlie spatiotemporal neural codes for odors. J Neurosci. 2010 Feb 10;30(6):1994-2006. doi: 10.1523/JNEUROSCI.5639-

09.2010. PMID: 20147528; PMCID: PMC2835415.

2. https://towardsdatascience.com/step-by-step-signal-processing-with-machine-learning-pca-ica-nmf-8de2f375c422

3. https://www.mathworks.com/help/images/discrete-cosine-transform.html

4. https://www.mathworks.com/help/signal/ug/hilbert-transform.html

5. https://www.mathworks.com/discovery/empirical-mode-decomposition.html

6. https://www.mathworks.com/help/signal/ref/stft.html

7. https://www.mathworks.com/help/stats/t-sne.html

6 Appendix- Other signal processing methods

6.1 Discrete cosine transform

The discrete cosine transform (DCT) represents an image as a sum of sinusoids of varying magnitudes and frequencies. The dct2 function computes the two-dimensional discrete cosine transform (DCT) of an image. The DCT has the property that, for a typical image, most of the visually significant information about the image is concentrated in just a few coefficients of the DCT. For this reason, the DCT is often used in image compression applications. For example, the DCT is at the heart of the international standard lossy image compression algorithm known as JPEG.

6.2 Hilbert transform

The Hilbert transform facilitates the formation of the analytic signal. The analytic signal is useful in the area of communications, particularly in bandpass signal processing. The toolbox function hilbert computes the Hilbert transform for a real input sequence x and returns a complex result of the same length, y = hilbert(x), where the real part of y is the original real data and the imaginary part is the actual Hilbert transform. y is sometimes called the analytic signal, in reference to the continuous-time analytic signal. A key property of the discrete-time analytic signal is that its Z-transform is 0 on the lower half of the unit circle. Many applications of the analytic signal are related to this property; for example, the analytic signal is useful in avoiding aliasing effects for bandpass sampling operations. The magnitude of the analytic signal is the complex envelope of the original signal. The Hilbert transform is related to the actual data by a 90-degree phase shift; sines become cosines and vice versa.

6.3 Empirical mode decomposition

Empirical mode decomposition (EMD) is a data-adaptive multiresolution technique to decompose a signal into physically meaningful components. EMD can be used to analyze non-linear and non-stationary signals by separating them into components at different resolutions. Some of the common applications of empirical mode decomposition are in the fields of bearing fault detection, biomedical data analysis, power signal analysis, and seismic signals. Empirical mode decomposition can be used to perform time-frequency analysis while remaining in the time domain. The components are in the same time scale as the original signal, which makes them easier to analyze. Unlike other multiresolution analysis (MRA) techniques such as wavelet analysis, empirical mode decomposition recursively extracts different resolutions from the data itself without the use of fixed functions or filters. Another way to explain EMD is to consider a signal as a fast oscillation superimposed on a slower one. After the fast oscillation is extracted, the EMD algorithm treats the remaining slower component as the new signal and again regards it as a fast oscillation superimposed on a slower one. The algorithm continues until some exit criterion is reached. The components in EMD are referred to as intrinsic mode functions (IMF).

6.4 Short-time Fourier transform

The short-time Fourier transform (STFT) is used to analyze how the frequency content of a nonstationary signal changes over time. The STFT of a signal is calculated by sliding an analysis window of length M over the signal and calculating the discrete Fourier transform of the windowed data. The window hops over the original signal at intervals of R samples.