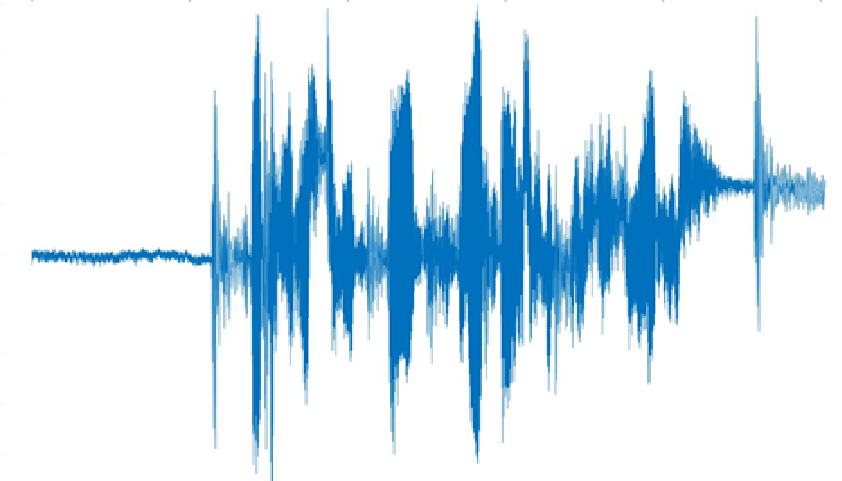
1. **Data Processing and Feature Extraction**

This chapter explains the process adopted for data extraction and processing such that an appropriate dataset is selected for constructing a Music Emotion Recognition (MER) model. The reference dataset adopted, MER Audio Traffic Data, contains sound samples classified into 4 quadrants that were adapted from Russell’s circumplex model of emotion. The dataset (N=900) has 225 samples of sound in each quadrant thereby making it a balanced dataset. The independent variable in this dataset is the emotion quadrant representation of sound signal. Since there are 4 quadrants, the independent variable is a multi-label variable with 4 factors viz. “Q1”, “Q2”, “Q3” and “Q4”. Each factor corresponds to a quadrant within the Russell’s Circumplex Model of emotion representation. The core objective of the data extraction and processing approach is to determine the best dataset that can aid modelling based on data sampling, data scaling, outlier detection and feature inclusion / exclusion.

* 1. **Intuition**

There are three dimensions to the characteristics of a sound – time dimension, frequency dimension and the strength or magnitude of the signal. A sound wave is typically perceived as a time variant signal and it is easy to visualize the time variability (on a coordinate plane x-axis represents the time and y-axis represents the magnitude). However, there’s a frequency axis that also varies with time (represented by z-axis in a coordinate plane) indicating that for a single unit of propagation of time along x-axis, there are several frequencies that are activated with varying degree of strength or magnitude. This characteristic makes the sound “time-variant” and “frequency-variant” as illustrated in the figure below (Fig number: TBD). This characteristic also differentiates a sound signal from a normal signal where the frequency component is comparatively constant.



Original Sound signal

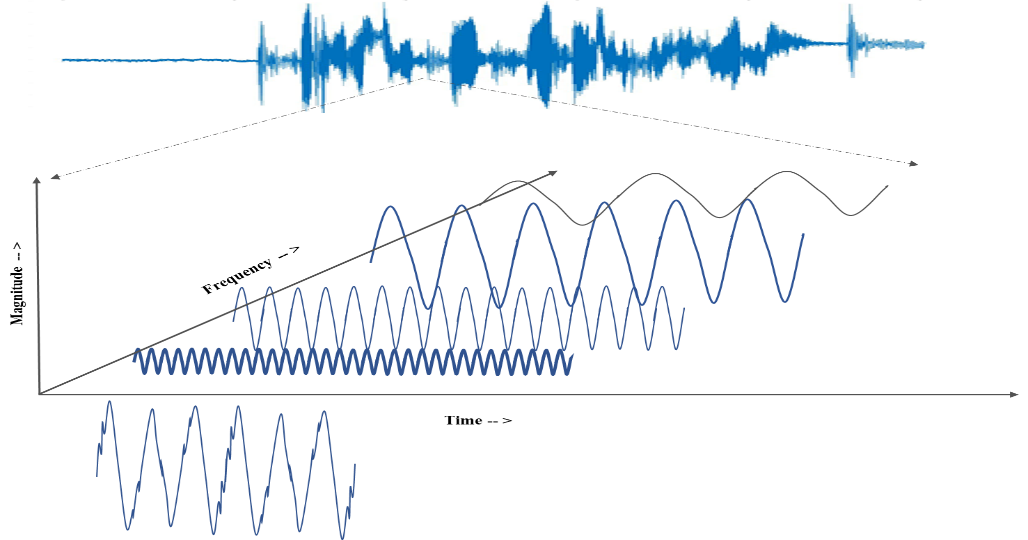


Illustration of a small window of sound signal into its frequency and time variant components (not to true scale)

In case of musical sounds, the presence of frequency components dramatically increases driven by the timbre of predominant sound (in case of monophonic sound) and each musical instrument in use (in case of polyphonic sound). This increase in frequency components imposes the need to extract appropriate spectral component from sound signal apart from its time variant or temporal characteristics. In addition, due to the abundance of data, appropriate sampling and outlier treatment should be applied on sound sample so that an abstract representation of sound is illustrated in the datasets that would be fed to modelling. Hence “influential dataset” selection strategy was adopted to arrive at an optimal dataset that could be used for modelling.

* 1. **Data Sampling & Transformation**

Data sampling & transformation is an important aspect associated with sound processing. For the purpose of this research, three broad data sampling strategies were implemented viz.

1. Each sound file in the dataset which is 30 seconds long, is considered as a single observation and features were extracted from this observation. This dataset is henceforth called as “base dataset”
2. Each sound file in the dataset is divided into multiple samples with each sample being 5 second long windows and with each window having a 1 second overlap with its predecessor. The features were extracted from every 5 second long window and inherited the class (emotion quadrant) of the overall sound sample
3. Each sound file in the dataset is divided into multiple samples with each sample being a window of 1 second duration and with each window having a 1/4 second overlap with its predecessor. The features were extracted from each of the 1 second long window and the label (emotion quadrant) of the overall sound sample as the class label

In addition, the whole dataset was transformed using Hilbert Transformation (HT) (ref:TBD) as it is robust to noise and offers an abstract representation of the sound. The aforementioned sampling strategy is applied on the raw sounds as well Hibert Transformed sounds thereby resulting in six datasets viz. base dataset - no HT, base dataset – HT, 5 seconds window dataset – no HT, 5 seconds window dataset – with HT, 1 second window dataset – no HT and 1 second window dataset – with HT.

* 1. **Feature Extraction**

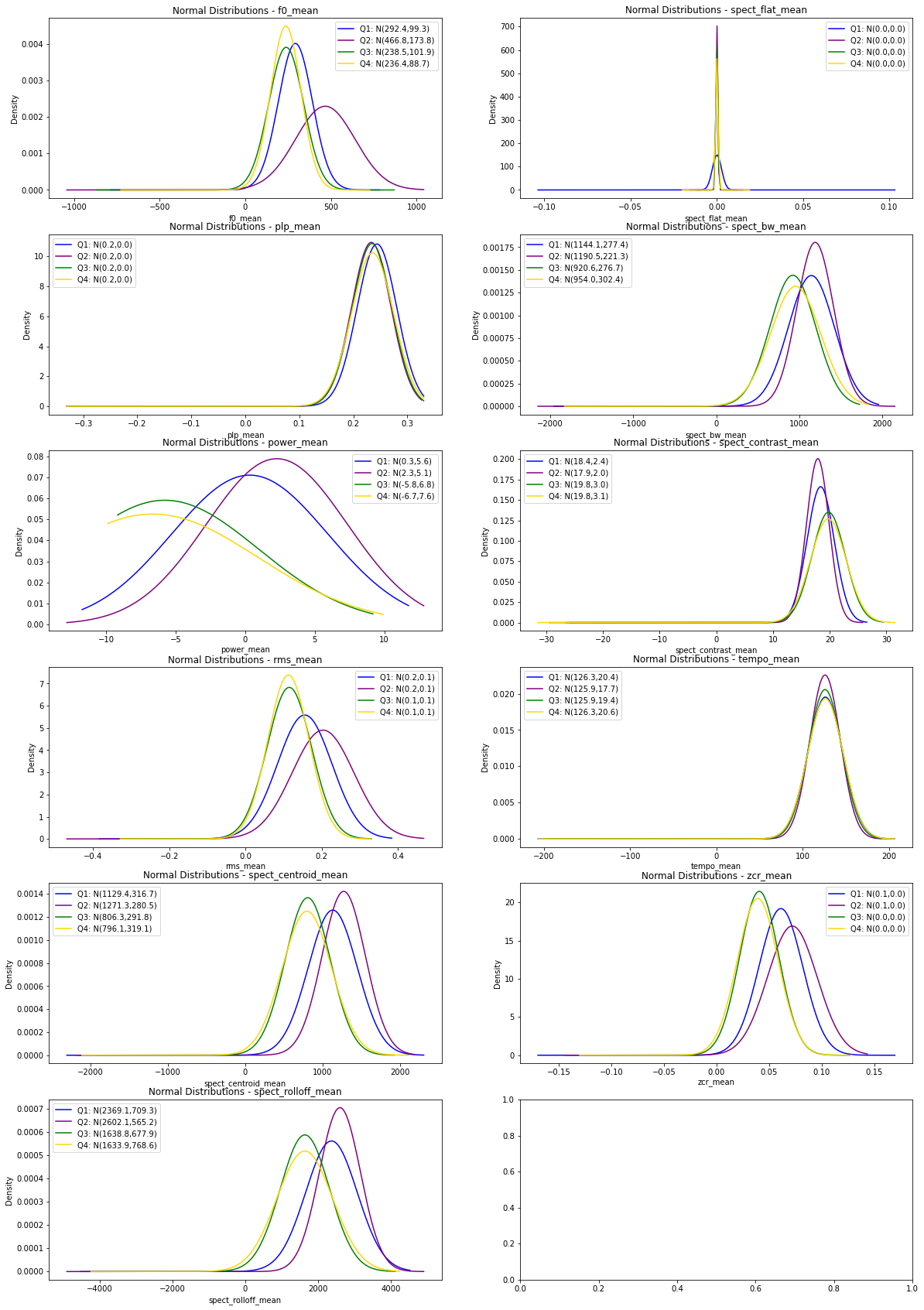
Based on this intuition, following fundamental temporal and spectral features were extracted from the musical sounds in the each of the six datasets generated.

* Fundamental Frequency (F0)
* RMS
* Spectral Centroid
* Spectral RollOff
* Spectral Flatness
* Spectral Bandwidth
* Spectral Contrast
* Zero Crossing Rate
* Tempo
* Predominant Local Pulse (PLP)
* Power
* 20 MFCC’s
* 64 Mel Frequencies
* Loudness

Key measures such as Mean, median, min, max, IQR and standard deviation, of each of the aforementioned features were extracted from each sound sample and the table below illustrates value ranges of features within base dataset.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **N = 900, Q1=225, Q2= 225, Q3 = 225, Q4 = 225** | | | | | | |
| **Feature** | **class** | **mean** | **median** | **max** | **sd** | **iqr** |
| Fundamental  Frequency (f0) | Q1 | 169.229 - 682.105 | 132.501 - 806.396 | 1050 - 1050 | 76.135 - 362.434 | 24.743 - 817.885 |
| Q2 | 178.455 - 878.958 | 130.473 - 1050.0 | 1050 - 1050 | 107.8 - 440.867 | 27.794 - 919.527 |
| Q3 | 136.67 - 582.762 | 130.473 - 615.892 | 1050 - 1050 | 44.726 - 382.083 | 0.0 - 770.725 |
| Q4 | 151.714 - 567.331 | 130.473 - 486.91 | 1050 - 1050 | 48.355 - 403.464 | 0.0 - 854.414 |
| Predominant  Local  Pulse (PLP) | Q1 | 0.097 - 0.277 | 0.0 - 0.033 | 1-Jan | 0.141 - 0.338 | 0.172 - 0.603 |
| Q2 | 0.044 - 0.28 | 0.0 - 0.013 | 1-Jan | 0.099 - 0.342 | 0.06 - 0.614 |
| Q3 | 0.098 - 0.264 | 0.0 - 0.04 | 1-Jan | 0.155 - 0.324 | 0.106 - 0.575 |
| Q4 | 0.076 - 0.28 | 0.0 - 0.016 | 1-Jan | 0.131 - 0.342 | 0.11 - 0.616 |
| Power | Q1 | -14.589 - 13.717 | -6.189 - 21.679 | 22.041 - 39.348 | 18.159 - 33.365 | 6.484 - 69.119 |
| Q2 | -10.758 - 14.389 | 0.699 - 22.839 | 15.578 - 39.126 | 14.07 - 29.502 | 9.01 - 45.922 |
| Q3 | -14.938 - 10.896 | -9.284 - 17.833 | 16.806 - 38.551 | 19.03 - 31.429 | 8.361 - 71.451 |
| Q4 | -21.873 - 10.813 | -19.215 - 21.422 | 14.559 - 39.492 | 17.155 - 30.082 | 8.846 - 70.988 |
| RMS | Q1 | 0.031 - 0.332 | 0.026 - 0.341 | 0.126 - 0.77 | 0.021 - 0.16 | 0.021 - 0.214 |
| Q2 | 0.022 - 0.399 | 0.021 - 0.43 | 0.058 - 0.689 | 0.008 - 0.15 | 0.009 - 0.246 |
| Q3 | 0.016 - 0.28 | 0.008 - 0.294 | 0.083 - 0.7 | 0.013 - 0.17 | 0.012 - 0.324 |
| Q4 | 0.017 - 0.243 | 0.017 - 0.235 | 0.048 - 0.75 | 0.009 - 0.142 | 0.011 - 0.203 |
| Spectral  Bandwidth | Q1 | 393.436 - 1786.716 | 393.844 - 1843.467 | 964.397 - 3193.479 | 68.176 - 480.811 | 54.672 - 636.891 |
| Q2 | 515.302 - 2015.781 | 491.474 - 2096.548 | 1312.649 - 3136.144 | 61.459 - 407.248 | 62.439 - 760.684 |
| Q3 | 291.586 - 1611.415 | 267.225 - 1643.87 | 1039.565 - 3255.02 | 51.257 - 507.864 | 51.458 - 742.657 |
| Q4 | 196.51 - 1722.461 | 190.308 - 1942.784 | 1177.24 - 3385.334 | 55.389 - 493.28 | 48.233 - 738.594 |
| Spectral  Centroid | Q1 | 373.015 - 1814.521 | 368.615 - 1773.51 | 1079.09 - 7653.906 | 65.577 - 945.845 | 73.341 - 1674.926 |
| Q2 | 434.031 - 1994.109 | 379.542 - 1996.843 | 1561.807 - 7621.4 | 102.138 - 766.088 | 120.463 - 1277.371 |
| Q3 | 315.243 - 1722.115 | 271.812 - 1675.192 | 1363.449 - 7693.994 | 77.2 - 733.496 | 97.239 - 882.259 |
| Q4 | 186.474 - 1924.253 | 183.155 - 2221.394 | 1051.875 - 7635.677 | 36.68 - 866.793 | 39.372 - 1666.57 |
| Spectral  Contrast | Q1 | 19.869 - 27.383 | 16.0 - 26.021 | 41.904 - 82.556 | 7.527 - 19.834 | 7.091 - 24.035 |
| Q2 | 18.816 - 28.167 | 13.967 - 27.934 | 48.891 - 82.558 | 8.592 - 20.738 | 5.564 - 23.014 |
| Q3 | 20.57 - 29.201 | 15.717 - 28.66 | 43.077 - 82.946 | 7.571 - 19.859 | 6.909 - 21.894 |
| Q4 | 20.339 - 30.835 | 17.174 - 30.246 | 44.606 - 82.765 | 5.993 - 19.573 | 7.63 - 21.456 |
| Spectral  Flatness | Q1 | 0.0 - 0.063 | 0.0 - 0.0 | 1.0 - 1.0 | 0.02 - 0.242 | 0.0 - 0.0 |
| Q2 | 0.0 - 0.012 | 0.0 - 0.0 | 1.0 - 1.0 | 0.02 - 0.109 | 0.0 - 0.0 |
| Q3 | 0.0 - 0.014 | 0.0 - 0.0 | 0.517 - 1.0 | 0.01 - 0.115 | 0.0 - 0.0 |
| Q4 | 0.001 - 0.007 | 0.0 - 0.0 | 1.0 - 1.0 | 0.021 - 0.082 | 0.0 - 0.0 |
| Spectral  Rolloff | Q1 | 660.864 - 3903.849 | 592.163 - 4080.542 | 2099.487 - 9420.776 | 161.71 - 1607.156 | 172.266 - 2640.509 |
| Q2 | 795.336 - 4479.259 | 742.895 - 4737.305 | 3445.312 - 8839.38 | 184.569 - 1262.966 | 204.565 - 2422.485 |
| Q3 | 500.531 - 3800.328 | 398.364 - 4037.476 | 3100.781 - 9087.012 | 189.785 - 1673.362 | 161.499 - 2842.383 |
| Q4 | 259.488 - 3796.873 | 258.398 - 4586.572 | 2487.085 - 9076.245 | 54.761 - 1512.542 | 43.066 - 2767.017 |
| Tempo | Q1 | 89.103 - 172.266 | 89.103 - 172.266 | 89.103 - 172.266 | 0 - 0 | 0 - 0 |
| Q2 | 89.103 - 191.406 | 89.103 - 191.406 | 89.103 - 191.406 | 0 - 0 | 0 - 0 |
| Q3 | 87.593 - 172.266 | 87.593 - 172.266 | 87.593 - 172.266 | 0 - 0 | 0 - 0 |
| Q4 | 82.031 - 184.57 | 82.031 - 184.57 | 82.031 - 184.57 | 0 - 0 | 0 - 0 |
| Zero Crossing  Rate (ZCR) | Q1 | 0.017 - 0.105 | 0.011 - 0.101 | 0.081 - 0.444 | 0.01 - 0.096 | 0.012 - 0.161 |
| Q2 | 0.023 - 0.134 | 0.017 - 0.13 | 0.077 - 0.443 | 0.01 - 0.085 | 0.014 - 0.127 |
| Q3 | 0.015 - 0.114 | 0.012 - 0.114 | 0.047 - 0.555 | 0.004 - 0.064 | 0.005 - 0.078 |
| Q4 | 0.01 - 0.107 | 0.007 - 0.098 | 0.047 - 0.438 | 0.004 - 0.068 | 0.006 - 0.101 |

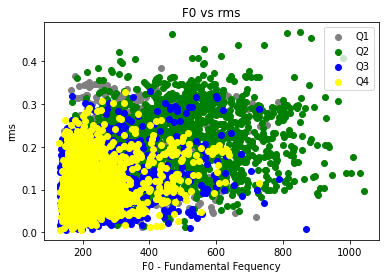
In order to examine this data further, normal distribution plots were elicited for each of the key features listed in the table above with the emotion class variable acting as the differentiator in each plot. To draw the normal distribution plot, the mean and standard deviation of mean value aggregation of the features were considered. These plots, as illustrated in the figure below, were intended to only for a visual inspection and not intended to determine the normality of the observations since the intended models, that would be used subsequently, are robust classifiers that a agnostic to underlying distribution of data. Hence the univariate normality tests were not performed on the aforementioned observations associated with aforementioned features.

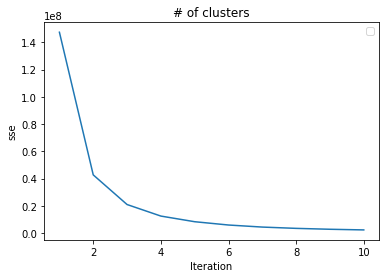


The normal distribution plot illustrates significant class overlap between the emotion quadrants Q1, Q3 and Q4. These quadrants represent Happy, Sad and Surprise emotions whereas quadrant Q2 represents Anger, Fear and Disgust emotions. As the dataset is annotated, the class overlap might indicate that through musically induced emotion one can differentiate between:

* + - Happy Vs Fear or Anger or Disgust
    - Sad Vs Fear or Anger or Disgust
    - Surprise Vs Fear or Anger or Disgust

However, when it comes to differentiating Happy Vs Sad emotion, there is significant bias and subjectivity involved as indicated by the class overlap. In order to further determine the actual number of differentiable classes in the dataset, k-means clustering method is used so as to determine optimal clusters in the dataset using elbow method. The plot below illustrates the outcome of k-means clustering.





As indicated by k-means clustering there a 3 groups or clusters present in the dataset. Hence a new class variable was defined to map these 3 groups to emotion quadrants, depending upon the impurity of original class variable in each cluster group, so that a clear separation could be achieved between quadrants Q2 vs the rest. This also reduces the number of levels in class variable from 4 to 3. Defining a new class variable will also help in determining the best approach during modelling as the outcomes with 4 level class variables (original emotion quadrant mapping) can be combined with the derived 3 level class variables (original emotion quadrant mapped to each cluster/group).

* 1. **Data Scaling & dressing**

As evident from the table above, the range of values differs for each feature, indicating a need for scaling the data. Hence a simple scaling strategy, as illustrated below, was adopted to scale and dress data.

* The null observations within each feature of dataset are imputed with 0
* Scaling range adopted is 0 to 1

The scaled dataset is then subjected to “dressing” – a process in while “irrelevant” features were eliminated from the dataset. The definition of “irrelevant” from the context of this research is as follows:

* A feature that contains a constant value
* A feature with a long-tailed distribution i.e., features with more than 95% observations being constant values

As a result of scaling and dressing, the number of datasets became 12 – six datasets from sampling (unscaled) and six datasets that were scaled.

* 1. **Outlier detection & treatment**

Presence of outliers in feature sets can always induce bias in model and can skew the model outcome. Hence it is important to treat the outliers appropriately to ensure that the model behavior is validated without any loss of critical information. The presence of outliers was evaluated from three perspectives – contextual anomalies, collective anomalies and point anomalies.

Technically, the dataset may contain two categories of outliers – musical and data. From a musical perspective, the outliers may include melodic fissure (unintended presence of a wrong note, in accordance with the key, in note sequence), tempo drag (increase / decrease of speed), power variation (impulsive stress in the note that can add a spike to the “attack” element of note – *explain the attach, sustain and decay of musical note as a diagram*) , sustain stretch (unintentional sustaining of a musical note for a larger duration than intended) etc. to name a few, arising at random moments within the musical propagation. The musical outliers were also considered as contextual anomalies. Since the context of data annotation is not defined and that the meta data explanation doesn’t cover aspects associated with aforementioned outliers, the musical outliers were not treated as a part of outlier treatment process.

The potential presence of collective anomalies can arise from several scenarios – intended or unintended deviation from the melodic structure for a specific period of time, embedding of elongated silence etc. Since these are also related to musical context and the annotation doesn’t define it clearly, such anomalies or outliers were also considered out of scope.

As a result of this scoping, subject to musical context, the only possible outlier treatment applicable was to treat the point anomalies that may arise due to presence of noise – either during the recording of sound or during the data transformation (sampling, digitization, music augmentation with effects such as distortion, delay, phaser etc.). The point anomaly / outlier detection strategy included identifying outlier observation in each feature after standardization using z-score and imputing any observation with z-score > 3 with mean value so that influential observations could be retained in the dataset and only extreme outliers are eliminated which in turn aids in controlling loss of information. The implementation of outlier detection strategy increased the number of datasets to 24 – 12 datasets without outlier treatment and 12 with outlier treatment.

* 1. **Feature inclusion / exclusion**

The presence of MFCC components and Mel Frequency components can lead to high correlation even though the boosting & bagging classification are theoretically agnostic to it. Hence to determine the impact of presence of autocorrelated components, separate datasets were generated without MFCC features in it. This separation resulted in number of datasets growing to 48 – 24 datasets with MFCC features and 24 datasets without MFCC features.

* 1. **Finding optimal dataset for modelling**

Each generated dataset is then used to build classification model using AdaBoost. Grid Search method is used to determine initial set of optimal parameters from following list of parameters and values: learning rate = [0.01,0.05,0.1,0.2,0.3] and n\_estimators: [100,200,500,800,1000,1500]. A repeated stratified K-fold (K=5, n\_repeats=1) is used as the cross-validation model with Grid Search with Accuracy was the metric to obtain optimal parameters. The outcome of the process to find optimal parameters of AdaBoost based on each dataset is illustrated in the table below.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **# classes** | **Sample Duration** | **N** | **MFCC included?** | **Hilbert Transformed?** | **Data Scaled?** | **Anomaly Treated?** | **# features** | **learning**  **rate** | **n estimators** | **Best Accuracy** |
| 4 | 30 seconds | 900 | No | No | No | No | 390 | 0.01 | 500 | 0.5367 |
| Yes | 510 | 0.3 | 800 | 0.5478 |
| Yes | No | 390 | 0.01 | 500 | 0.5367 |
| Yes | 510 | 0.3 | 800 | 0.5478 |
| Yes | No | No | 385 | 0.01 | 1500 | 0.5300 |
| Yes | 505 | 0.2 | 800 | 0.5556 |
| Yes | No | 385 | 0.01 | 1500 | 0.5300 |
| Yes | 505 | 0.2 | 800 | 0.5556 |
| Yes | No | No | No | 510 | 0.05 | 200 | 0.5411 |
| Yes | 510 | 0.1 | 200 | 0.5433 |
| Yes | No | 510 | 0.05 | 200 | 0.5411 |
| Yes | 510 | 0.1 | 200 | 0.5433 |
| Yes | No | No | 505 | 0.1 | 100 | 0.5422 |
| Yes | 505 | 0.3 | 1000 | 0.5611 |
| Yes | No | 505 | 0.1 | 100 | 0.5422 |
| Yes | 505 | 0.2 | 800 | 0.5556 |
| 5 seconds | 6287 | No | No | No | No | 433 | 0.3 | 1500 | 0.5998 |
| Yes | 553 | 0.1 | 1500 | 0.6331 |
| Yes | No | 433 | 0.1 | 1500 | 0.5998 |
| Yes | 553 | 0.1 | 1500 | 0.6331 |
| Yes | No | No | 427 | 0.2 | 1000 | 0.5769 |
| Yes | 547 | 0.1 | 1500 | 0.6162 |
| Yes | No | 427 | 0.2 | 1000 | 0.5779 |
| Yes | 547 | 0.1 | 1500 | 0.6162 |
| Yes | No | No | No | 553 | 0.2 | 1000 | 0.6095 |
| Yes | 553 | 0.1 | 1500 | 0.6331 |
| Yes | No | 553 | 0.3 | 500 | 0.6119 |
| Yes | 553 | 0.1 | 1500 | 0.6331 |
| Yes | No | No | 547 | 0.3 | 800 | 0.5958 |
| Yes | 547 | 0.1 | 1500 | 0.6162 |
| Yes | No | 547 | 0.3 | 800 | 0.5973 |
| Yes | 547 | 0.1 | 1500 | 0.6162 |

Based on the outcome following inferences can be made:

* On average, the overall accuracy increases while determining optimal parameters for AdaBoost model, when the underlying dataset is augmented (with 5 second window samples) in comparison with non-augmented dataset (30 second samples). The increase was significant; χ2 (1) = 23.40, *p* <0.001, *d*=4.58, as illustrated by Kruskal-Wallis rank sum test. This indicates that data augmentation may be performed while drawing sound sampled for constructing MER models
* The accuracy marginally increases while determining optimal parameters for AdaBoost model when the underlying dataset used is outlier treated. Kruskal-Wallis rank sum test suggests that this increase is significant though the effect size is low; χ2 (1) = 5.85, *p* =0.016, *d*=0.6. Based on the significance observed, the outlier treatment, as indicated in section ***tbd*** above, may be performed to obtain a better outcome while constructing MER model
* Kruskal-Wallis rank sum test suggests that inclusion of MFCC features have no effect on either accuracy or optimal parameter determination for AdaBoost model χ2 (1) = 0.173, *p* =0.678, indicating that the these features may be excluded from the corpus to be used for constructing MER model
* Kruskal-Wallis rank sum test suggests that performing Hilbert Transform of the sound signal have no effect on either accuracy or optimal parameter determination for AdaBoost model; χ2 (1) = 0.091, *p* =0.762, indicating that the sound signal may not be Hilbert Transformed for constructing MER model
* Kruskal-Wallis rank sum test also suggests that data scaling, as illustrated in section ***tbd*** above no effect on either accuracy or optimal parameter determination for AdaBoost model; χ2 (1) = 0, *p* =1, indicating that the features included in the dataset may not be scaled while constructing MER model

Based on the inferences the MFCC features were removed from the raw dataset. In addition, the dataset was outlier treated and data augmented before deriving the final dataset for constructing the MER model.

* 1. **Validation dataset**

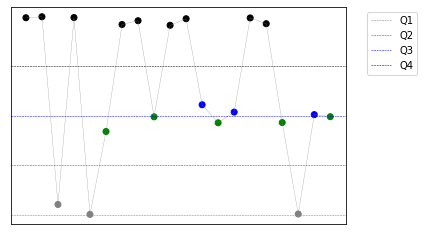
The validation dataset contains songs (downloaded from raaga.com) sung by various artists and composed to Carnatic ragas Bhairavi, Kalyani, Kedaragaula, Khamboji, Mayamalavagaula, Mohanam, Shankarabharanam and Todi. The features were extracted from these sound samples would correspond to the outcome of the dataset analysis, as described in section (TBD) above. This feature set would act as the validation dataset that would be used by the model to draw inferences. (*available at: D:\PhD Program\Phd - Data Science\Research Project\Carnatic Songs\Vocals*)

* 1. **Emotion Plot**

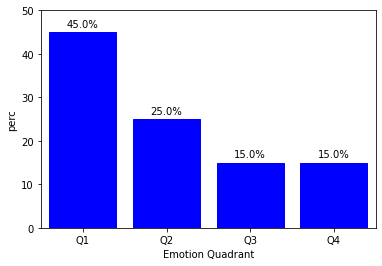
As evident from the analysis explained above, it may be appropriate and accurate to determine the emotion classification of a sound signal with a shorter duration (5 second) than with a long duration (30 seconds). As a result, it may not be appropriate to take a whole song and determine the emotion classification of it since the duration of the song would be more than 5 seconds. Hence an approach is devised to determine to not only capture overall discreet emotion associated with a song but also to evaluate variations in emotions within a song. Following are the steps involved in the proposed approach

* + - Define a window with a duration of 5 seconds
    - Slide this window across the song so as the slice the song, but with 1 second overlap with the previous slice
    - Take each slice of the song, extract relevant features and feed it as an input to the MER model
    - Determine emotion classification of the slice and plot in on a “emotion quadrant plot”
    - Repeat the process for all slices
    - Determine the frequency of each emotion class so that the class with highest frequency can be assigned as the discreet emotion class for the entire sound

The emotion quadrant plot is a simple x-y scatter plot with four bands representing each emotion is plotted. Here’s an example of the plot simulating the model outcome of a sound with ~100 seconds duration.



In this plot it can be observed that more slices are classified as Q1 where are some slices are classified in other quadrants. Hence this plot can reflect the trend of potential emotional variation observed within the song and the bar plot below, which is a frequency distribution of emotion classes within a song, illustrates the discreet representation & quantification of emotion within a given song.



*Note: the plots above are based on simulated data and is illustrated here to explain the intuition associated with deciphering model outcome*

The combination of both plots is henceforth called as “Emotion Plot” and would be relied to evaluate the emotion trends of raga Mayamalavagowla to draw inferences associated with the research question. In addition, other ragas, as mentioned in section 4.8: Validation set, would also be evaluated using the Emotion Plot.

* 1. **Conclusion**