Columbia University

Project Report

COMS 6901

Correlation of acoustic features of speakers with Audience Engagement

Author:
Animesh Sharma

Supervisor:
Dr. John Kender

January 1, 2018

Abstract

In this work, we investigate the extent to which the acoustic features of speakers are correlated to the audience engagement. The number of eyeblinks in a particular time period was taken as a measure of engagement of audience and the acoustic feature set of size 6373 was obtained for two different TEDx videos- Simon and Carol, with the help of OpenSmile software. Then eigen-analysis and mutual info regression based correlation analysis were used for feature selection in case of audio feature set. In order to determine the extent to which the acoustic features are correlated with the attention signal, two learning algorithms- Multi-layer Perceptron (MLP) and Support Vector Regression (SVR), were used to predict the attention signal with the acoustic feature set being used as input to the network. The best R^2 score of 0.55 was obtained for the correlation based analysis of Simon video with Support Vector Regression and acoustic feature set of size 200. Then the anomalies in audience, which were previously being handled through Z-score, were removed and all the experiments were performed again. This time the best result was 0.57 for Simon video with correlation based analysis and Support Vector Regression. The audio features most correlated with eyeblink data were also listed for both cases with feature selection based on correlation analysis.

Contents

1 Introduction

This project is based on the work done by Zhang et al. [1]. in which the authors propose that certain speaker gestures can convey important information regarding audience engagement. Two TEDx videos, labeled Simon and Carol, were used for this project and it was studied whether certain acoustic features have more influence on the audience engagement. This work is useful in determining whether audience engagement can be improved with the help of the observed acoustic features. The acoustic features most correlated with the audience engagement can help instructors and speakers in further engaging their audience. The eyeblink data was used as a measure of the audience engagement based on the premise that the subject tends to pay more attention when the number of eyeblinks are less for that time period. There were 28 subjects for both the videos and the number of eyeblinks for a 5 second window were obtained for the audience as a whole by taking the mean for that time period. The acoustic feature dataset of size 6373 was obtained using OpenSmile [2]. software and this dataset was then used as input to the learning algorithms. This acoustic feature set, known as ComParE feature set [3], was given by Eybet et al [4] and they are functionals of acoustic low level descriptors (LLDs). Then feature selection was done for the audio feature set so as to evaluate the performance of the learning algorithms only on the relevant features and to filter out the noise. The learning algorithms were used to determine whether the attention signal can be predicted with acoustic signal as input. The performance of the learning algorithm give a sense of the importance of audio features in determing the audience engagement. Furthermore, the important audio features determined by correlating them with attention signal can also be listed so that it can be known which audio features out of the feature set of size 6373 influence the attention signal most. Predicting the attention signal is a regression problem and two different learning algorithms were used for the same. One important issue with this task was handling the anomalies present in the eyeblink dataset of audience. Some of the eyeblink recordings of the subjects were corrupted in the sense that there were two many eyeblinks in certain 5 second windows. These anomalies were handled by taking Z-scores of the recordings in first case and afterwards removing them in the second. All the steps were repeated for both the cases.

2 Problem Formulation

Two different datasets- eyeblink dataset and acoustic feature dataset, were prepared for this work and then subsequent analysis was done for determining how much one is correlated with the other.The whole project pipeline has 5 subsections and these were executed in-order to obtain the final result. The following contributions were made in this work:

1. The eyeblink data was cleaned and pre-processed to get an attention signal for the audience. The experiments were conducted first for all the 28 subjects in audience and then the anomalies were removed and the experiments were repeated for 17 subjects. The outlier subjects were removed on the criteria that for these subjects, the number of eyeblinks in a 5 second window was too high for a human being.

2. The acoustic feature dataset was prepared for the two TEDx videos for 5 second windows using OpenSmile software. The audio was stripped off from the video file and 5 second windows were obtained using ffmpeg wrapper [5].

3. The acoustic feature set of size 6373 was reduced to sizes 10, 20, 50, 100 and 200 using eigenanalysis and mutual information based correlation analysis. The scikit-learn library in Python [6] was used for this feature selection. The reduced feature set in eigenanalysis was further preprocessed using Z-score method.

4. Then it was determined whether its possible to predict the attention of the audience with reduced acoustic feature set. Two different learning algorithms were used - Neural Networks and Support Vector Regression. The scikit-learn was again used here and the network settings for both the methods are defined in subsequent sections.

5. The different acoustic features obtained for both the full audience and the reduced audience were also listed and the meaning of these features was obtained so that it can be determined which characteristics of a person's speech are helpful in capturing attention of the audience.

3 Proposed Method

The following subsections define the different submodules executed for obtaining the final result given in this work.

3.1 Eyeblink data cleaning and preprocessing

The raw data was in xls format and there were many redundant columns in it. The format was converted to csv. Multiple scripts in python were written for different preprocessing tasks. Only the follwing columns were retained- `CURRENT_FIX_LABEL` column which gives the fixation timestamp and the `NEXT_SAC_CONTAINS_BLINK` column which gives boolean values as to whether this section contains eyeblink or not. The data of the fixation column was converted to integer values and then the number of eyeblinks in a 5 second window was obtained. The Z-score values were obtained so as to normalize the eyeblink data with respect to each subject individually. Finally the mean of the eyeblink data for the audience was taken for each 5 second window.

3.2 Handling Outliers in Eyeblink dataset

For the first experiment, just the Z-score values were obtained for all the subjects. It was done so as to diffuse the effect of the outliers. There were 28 subjects for both the Simon and Carol video.

For the second experiment, the anomalies were removed for both the Simon and Carol videos. For the Carol video, the following subjects were removed with the given reasons:

- `jc1_carol`: consistent eyeblinks of 6 or 5
- `jhh1_carol`, `msb1_carol`: consistently high eyeblinks of 10
- `jy1_carol`, `pp1_carol`, `rn1_carol`: no eyeblinks at all (0 or 1)
- `pc1_carol`: lots of 10 at the end of the video
- `sa1_carol`, `sd2_carol`, `sw1_carol`, `ys1_carol`: consistently high eyeblinks

For the Simon video, the following subjects were removed with the given reasons:

- `jg1_simon`, `pp1_simon`, `rn1_simon`, `jy1_simon`: no eyeblinks at all (0 or 1)
- `sa1_simon`, `jc1_simon`, `sd2_simon`, `klt1_simon`, `azz1_simon`, `ys1_simon`, `jhh1_simon`: consistently high eyeblinks

3.3 Acoustic Feature Set

The audio was extracted from the video files in *mp3* format and then it was converted to *wav* format using ffmpeg wrapper. Then a script in *Ruby* was written to obtain 5 second windows for both the TEDx files and the offset was taken as 1 second here. Finally the ComParE audio feature set of size 6373 was obtained using OpenSmile and the output format was *csv*.

The length of the Carol video was 624 seconds and therefore 620 fragments were obtained for this file. Similarly, 714 fragments were obtained for the Simon video with length 718 seconds.

3.4 Feature Selection for Acoustic Feature Set

Two different methods were used for reducing the size of the ComParE dataset obtained using OpenSmile. First the size was reduced using Eigenanalysis. The *sklearn* package present in *Python* was used here to reduce the dataset size to 10, 20, 50, 100 and 200. Then the values were normalized using Z-score method.

For the second method, correlation with the eyeblink data was used to select only n features out of the 6373 available so that the selected features have highest correlation values. The method used here to calculate the correlation values was *mutual info regression* present in *sklearn* package. The reduced dataset sizes were again 10, 20, 50, 100 and 200 and the predefined method *SelectKBest* available in *scikitlearn* was used.

3.5 Neural Networks

A Multi-layer perceptron is a network of simple neurons called Perceptron. Each neuron computes the output which is some linear or usually a non-linear function of the linear combination of the input and the weights including the

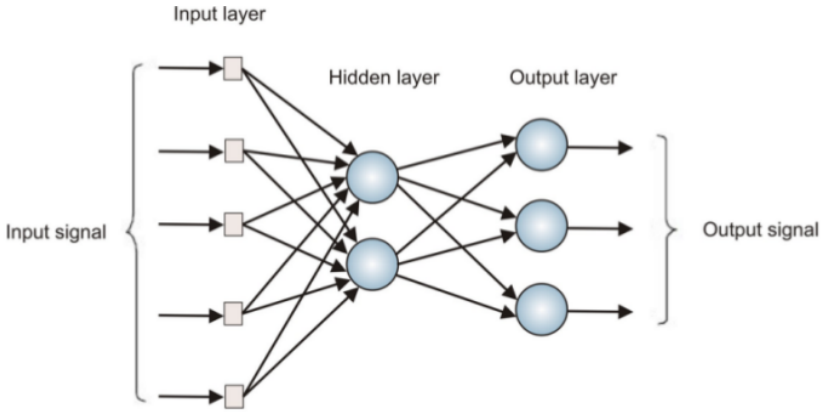

Figure 1: Structure of a Multi-layer Perceptron [7].

bias. A simple neuron has mapping limitations and thus a number or layers with a large number of neurons are used. The input neuron layer represents each instance which is fed to the hidden layer. The hidden layer computes the output depending on the input it receives, the corresponding weights and the activation function which is then fed to the output layer [7]. Thus the input signal propagates the network layer by layer. The number of the layers depends on the structure of the network. Figure 1 shows the structure of a multi-layer perceptron.

Multilayer perceptron with single layer can solve a wide variety of problems. The weights can initially be assumed randomly and then the network can be trained by backpropagation [8]. The implementation given in *scikitlearn* was used and the hyper-parameter values are given in Table 1.

3.6 Support Vector Regression

Support Vector Machine can be applied not only to the classification problems but also to the regression problems. Similar to classification, the basic aim is to optimize the generalization bounds for regression [9]. It uses an epsilon intensive loss function which ignores errors that are situated within the certain distance of the true value. Figure 2 below shows an example of one-dimensional linear regression function with epsilon intensive band. The variables measure the cost of the errors on the training points. The error is zero for all points inside the epsilon band. SV algorithm can be made nonlinear by simply pre-processing the training patterns with kernel functions.

hyper-parameter	values
Train-Test split	90:10
Hidden-layer dimensions	500, 100
Activation Function	Relu
Optimizer	Adam
Regularization Coefficient	0.0001
Batch Size	200
Learning Rate	0.001 (constant)
Maximum Iterations	200
Tolerance	1e-4

Table 1: Hyper-parameters for MLP setup.

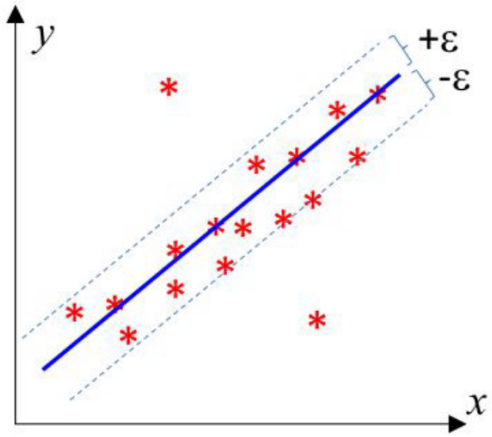

Figure 2: Linear regression function with epsilon insensitive band [9].

In the nonlinear setting, the optimization problem corresponds to finding the flattest function in feature space, not in input space. The implementation given in *scikitlearn* was used and the hyper-parameter values are given in Table 2.

hyper-parameter	values
Train-Test split	90:10
C	1.0
Epsilon	0.1
Kernel	RBF
Tolerance	1e-3

Table 2: Hyper-parameters for SVR setup.

4 Evaluation and Analysis

4.1 Datasets

Two TEDx videos were used of lengths 624 seconds and 719 seconds. Then acoustic feature datasets obtained were of dimensions 6373x620 and 6373x714. Feature selection reduced the dataset size to nx620 and nx714 where n was a hyper-parameter. The eyeblink data had dimensions 620x1 and 714x1.

4.2 Evaluation Metric

The R^2 score was used as a performance metric for the learning algorithms used in the last experiment. The coefficient R^2 is defined as:

$$R^2 = (1 - \frac{u}{v})$$

, where u is the residual sum of squares and v is the total sum of squares. The best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of target, disregarding the input features, would get a R^2 score of 0.0.

4.3 Results

The R^2 results for both the MLP and SVR setups were obtained in the first four experiments mentioned below. Four different experiments were conducted with the feature selection and learning methods. In the fifth experiment, the relevant acoustic features were determined. Since only correlation analysis gave good results, relevant features were determined for this feature selection.

$R^2 scores$		
Dataset Size	MLP	SVR
10	-0.13	-0.006
20	0.12	0.08
50	-0.09	0.20
100	-0.90	0.16
200	-1.52	0.24

Table 3: Results for Carol Video.

$R^2 scores$		
Dataset Size	MLP	SVR
10	-0.18	-0.12
20	0.29	0.16
50	0.03	0.13
100	-0.30	0.19
200	-0.28	0.25

Table 4: Results for Simon Video.

4.3.1 Full Audience and Feature Selection with Eigen-analysis

The follwing R^2 results were obtained for given reduced acoustic dataset sizes. Table 3 gives results for Carol TEDx video and Table 4 for Simon TEDx video. Figure 3 shows the plot for the best result obtained here.

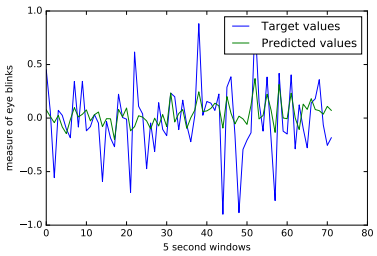

Figure 3: Plot for Simon Video with SVR and size 200.

$R^2 scores$		
Dataset Size	MLP	SVR
10	0.23	0.11
20	0.47	0.44
50	0.38	0.43
100	0.26	0.48
200	0.19	0.44

Table 5: Results for Carol Video.

$R^2 scores$		
Dataset Size	MLP	SVR
10	0.31	0.21
20	0.48	0.40
50	0.54	0.45
100	0.48	0.52
200	0.45	0.55

Table 6: Results for Simon Video.

4.3.2 Full Audience and Feature Selection with Correlation-analysis

The follwing R^2 results were obtained for given reduced acoustic dataset sizes. Table 5 gives results for Carol TEDx video and Table 6 for Simon TEDx video. Figure 4 shows the plot for the best result obtained here.

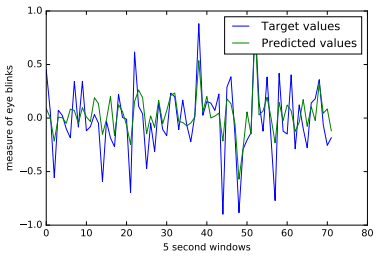

Figure 4: Plot for Simon Video with SVR and size 200.

$R^2 scores$		
Dataset Size	MLP	SVR
10	-0.52	0.09
20	-0.03	0.28
50	-0.008	0.31
100	0.03	0.32
200	-0.68	0.29

Table 7: Results for Carol Video.

$R^2 scores$		
Dataset Size	MLP	SVR
10	-0.32	-0.114
20	0.12	0.095
50	0.26	0.14
100	0.20	0.17
200	-0.17	0.16

Table 8: Results for Simon Video.

4.3.3 Reduced Audience and Feature Selection with Eigen-analysis

The follwing R^2 results were obtained for given reduced acoustic dataset sizes. Table 7 gives results for Carol TEDx video and Table 8 for Simon TEDx video. Figure 5 shows the plot for the best result obtained here.

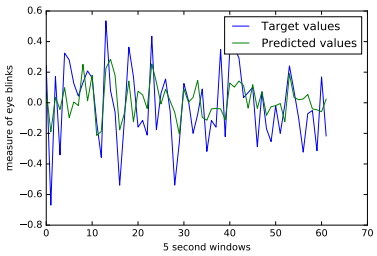

Figure 5: Plot for Carol Video with SVR and size 100.

$R^2 scores$		
Dataset Size	MLP	SVR
10	0.004	0.14
20	0.34	0.40
50	0.36	0.48
100	0.39	0.51
200	0.20	0.53

Table 9: Results for Carol Video.

$R^2 scores$		
Dataset Size	MLP	SVR
10	0.27	0.34
20	0.51	0.43
50	0.47	0.52
100	0.48	0.53
200	0.51	0.57

Table 10: Results for Simon Video.

4.3.4 Reduced Audience and Feature Selection with Correlation-analysis

The follwing R^2 results were obtained for given reduced acoustic dataset sizes. Table 9 gives results for Carol TEDx video and Table 10 for Simon TEDx video. Figure 6 shows the plot for the best result obtained here.

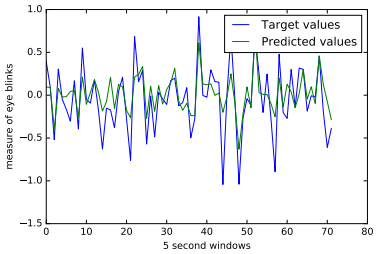

Figure 6: Plot for Simon Video with SVR and size 200.

4.3.5 Relevant Features with respect to Attention

Only the correlation-analysis gave good results. Therefore the relevant features obtained through this feature selection were determined for both the full audience case and reduced one.

Top 10 features for Carol video with Full Audience:

- `audspec_lengthL1norm_sma_stddev`:standard deviation of magnitude of the L1 norm of auditory spectrum and it was smoothed using an averaging filter with window length 3.

- `audSpec_Rfilt_sma[0]_range`:Rfilt means Relative Spectral Transform (RASTA)-style filtered. This means range of 0th Rfilt of auditory spectrum and it was smoothed using an averaging filter with window length 3.

- `audSpec_Rfilt_sma[19]_percentile1.0`:Rfilt means Relative Spectral Transform (RASTA)-style filtered. Percentile1.0 is the outlier-robust minimum value of the contour, represented by the 1 percentile. This means percentile1.0 of 19th Rfilt of auditory spectrum and it was smoothed using an averaging filter with window length 3.

- `audSpec_Rfilt_sma[24]_range`:Rfilt means Relative Spectral Transform (RASTA)-style filtered. This means range of 24th Rfilt of auditory spectrum and it was smoothed using an averaging filter with window length 3.

- `audSpec_Rfilt_sma[25]_percentile1.0`:Rfilt means Relative Spectral Transform (RASTA)-style filtered. Percentile1.0 is the outlier-robust minimum value of the contour, represented by the 1 percentile. This means percentile1.0 of 25th Rfilt of auditory spectrum and it was smoothed using an averaging filter with window length 3.

- `pcm_fftMag_spectralRollOff75.0_sma_leftctime`: The $75 * 100$ percent spectral roll-off point is determined as the frequency below which $75 * 100$ percent of the total signal energy fall. This means time during which the spectralRollOff75.0 of fast Fourier transformed pulse-code modulation has left curvature and it was smoothed using an averaging filter with window length 3.

- mfcc_sma[2]_range:range of 2nd mel-frequency cepstral coefficient and it was smoothed using an averaging filter with window length 3.
- mfcc_sma[4]_percentile1.0:Percentile1.0 is the outlier-robust minimum value of the contour, represented by the 1 percentile. This means percentile1.0 of 4th mel-frequency cepstral coefficient and it was smoothed using an averaging filter with window length 3.
- mfcc_sma[8]_percentile99.0:Percentile99.0 is the outlier-robust maximum value of the contour, represented by the 99 percentile. This means percentile99.0 of 8th mel-frequency cepstral coefficient and it was smoothed using an averaging filter with window length 3.
- mfcc_sma_de[4]_range:range of 4th mel-frequency cepstral coefficient and it was smoothed using an averaging filter with window length 3.

Top 10 features for Simon video with Full Audience:

- pcm_zcr_sma_percentile99.0:Percentile99.0 is the outlier-robust maximum value of the contour, represented by the 99 percentile. This means percentile99.0 of zero-crossing rate of time signal and it was smoothed using an averaging filter with window length 3.
- audspec_lengthL1norm_sma_de_range:range of magnitude of the L1 norm of auditory spectrum and it was smoothed using an averaging filter with window length 3.
- audSpec_Rfilt_sma[9]_range:Rfilt means Relative Spectral Transform (RASTA)-style filtered. This means range of 9th Rfilt of auditory spectrum and it was smoothed using an averaging filter with window length 3.
- audSpec_Rfilt_sma[15]_range:Rfilt means Relative Spectral Transform (RASTA)-style filtered. This means range of 15th Rfilt of auditory spectrum and it was smoothed using an averaging filter with window length 3.
- audSpec_Rfilt_sma[16]_pctlrange0-1:Rfilt means Relative Spectral Transform (RASTA)-style filtered. This means 0-1 inter-percentile range of 16th Rfilt of auditory spectrum and it was smoothed using an averaging filter with window length 3.

- `pcm_fftMag_spectralRollOff25.0_sma_range`: The $25 * 100$ percent spectral roll-off point is determined as the frequency below which 25 $*$ 100 percent of the total signal energy fall. This means range of spectralRollOff25.0 of fast Fourier transformed pulse-code modulation and it was smoothed using an averaging filter with window length 3.

- `pcm_fftMag_spectralHarmonicity_sma_range`:A Harmonicity object represents the degree of acoustic periodicity, also called Harmonics-to-Noise Ratio (HNR). This means range of spectral harmonicity of fast Fourier transformed pulse-code modulation and it was smoothed using an averaging filter with window length 3.

- `audSpec_Rfilt_sma_de[8]_range`:Rfilt means Relative Spectral Transform (RASTA)-style filtered. This means range of 8th Rfilt of auditory spectrum and it was smoothed using an averaging filter with window length 3.

- `pcm_fftMag_spectralSlope_sma_peakRangeAbs`:absolute value of range of all values for spectral slope of fast Fourier transformed pulse-code modulation and it was smoothed using an averaging filter with window length 3.

- `audspec_lengthL1norm_sma_de_peakRangeAbs`:absolute value of range of all values for magnitude of the L1 norm of auditory spectrum and it was smoothed using an averaging filter with window length 3.

Top 10 features for Carol video with Reduced Audience:

- `audspec_lengthL1norm_sma_quartile2`:$50th$ percentile of magnitude of the L1 norm of auditory spectrum and it was smoothed using an averaging filter with window length 3.

- `pcm_RMSenergy_sma_percentile99.0`:Percentile99.0 is the outlier-robust maximum value of the contour, represented by the 99 percentile. This means percentile99.0 of root-mean-square signal frame energy and it was smoothed using an averaging filter with window length 3.

- `pcm_RMSenergy_sma_pctlrange0-1`:This means 0-1 inter-percentile range of root-mean-square signal frame energy and it was smoothed using an averaging filter with window length 3.

- `audSpec_Rfilt_sma[20]_percentile1.0`:Rfilt means Relative Spectral Transform (RASTA)-style filtered. Percentile1.0 is the outlier-robust minimum value of the contour, represented by the 1 percentile. This means percentile1.0 of 20th Rfilt of auditory spectrum and it was smoothed using an averaging filter with window length 3.

- `mfcc_sma[1]_percentile99.0`:Percentile99.0 is the outlier-robust maximum value of the contour, represented by the 99 percentile. This means percentile99.0 of 1st mel-frequency cepstral coefficient and it was smoothed using an averaging filter with window length 3.

- `audSpec_Rfilt_sma_de[23]_percentile99.0`:Rfilt means Relative Spectral Transform (RASTA)-style filtered. Percentile99.0 is the outlier-robust maximum value of the contour, represented by the 99 percentile. This means percentile99.0 of 25th Rfilt of auditory spectrum and it was smoothed using an averaging filter with window length 3.

- `audSpec_Rfilt_sma[6]_meanRisingSlope`:meanRisingSlope is the mean of rising slopes, i.e. the slopes connecting a valley with the following peak. Rfilt means Relative Spectral Transform (RASTA)-style filtered. This means meanRisingSlope of 6th Rfilt of auditory spectrum and it was smoothed using an averaging filter with window length 3.

- `pcm_fftMag_spectralRollOff50.0_sma_minRangeRel`:The $50 * 100$ percent spectral roll-off point is determined as the frequency below which $50 * 100$ percent of the total signal energy fall. This means minimum value of relative range of spectralRollOff50.0 of fast Fourier transformed pulse-code modulation and it was smoothed using an averaging filter with window length 3.

- `audSpec_Rfilt_sma_de[21]_peakMeanAbs`:Rfilt means Relative Spectral Transform (RASTA)-style filtered. This means absolute value of mean of all values for 21th Rfilt of auditory spectrum and it was smoothed using an averaging filter with window length 3.

- `audSpec_Rfilt_sma_de[21]_peakMeanMeanDist`:Rfilt means Relative Spectral Transform (RASTA)-style filtered. This means mean of distance between peaks for 21th Rfilt of auditory spectrum and it was smoothed using an averaging filter with window length 3.

Top 10 features for Simon video with Reduced Audience:

- `audSpec_Rfilt_sma[13]_range`:Rfilt means Relative Spectral Transform (RASTA)-style filtered. This means range of 13th Rfilt of auditory spectrum and it was smoothed using an averaging filter with window length 3.

- `audSpec_Rfilt_sma[23]_range`:Rfilt means Relative Spectral Transform (RASTA)-style filtered. This means range of 23th Rfilt of auditory spectrum and it was smoothed using an averaging filter with window length 3.

- `pcm_fftMag_spectralRollOff25.0_sma_percentile99.0`:Percentile99.0 is the outlier-robust maximum value of the contour, represented by the 99 percentile. The $25 * 100$ percent spectral roll-off point is determined as the frequency below which $25 * 100$ percent of the total signal energy fall. This means percentile99.0 of spectralRollOff25.0 of fast Fourier transformed pulse-code modulation and it was smoothed using an averaging filter with window length 3.

- `pcm_fftMag_spectralFlux_sma_range`:range of spectral flux of fast Fourier transformed pulse-code modulation and it was smoothed using an averaging filter with window length 3.

- `pcm_fftMag_spectralSkewness_sma_range`:Spectral skewness is a measure of the asymmetry of the spectral distribution around its centroid. This means range of spectral skewness of fast Fourier transformed pulse-code modulation and it was smoothed using an averaging filter with window length 3.

- `pcm_fftMag_psySharpness_sma_percentile1.0`:psySharpness is related to how much a sound's spectrum is in the high end. Percentile1.0 is the outlier-robust minimum value of the contour, represented by the 1 percentile. This means percentile1.0 of psySharpness of fast Fourier transformed pulse-code modulation and it was smoothed using an averaging filter with window length 3.

- `mfcc_sma[6]_percentile1.0`:Percentile1.0 is the outlier-robust minimum value of the contour, represented by the 1 percentile. This means percentile1.0 of 6th mel-frequency cepstral coefficient and it was smoothed using an averaging filter with window length 3.

- `audSpec_Rfilt_sma_de[20]_range`:Rfilt means Relative Spectral Transform (RASTA)-style filtered. This means range of 20th Rfilt of auditory spectrum and it was smoothed using an averaging filter with window length 3.
- `pcm_fftMag_spectralFlux_sma_de_range`:delta of range of spectral flux of fast Fourier transformed pulse-code modulation and it was smoothed using an averaging filter with window length 3.
- `audSpec_Rfilt_sma_de[5]_stddevRisingSlope`:stddevRisingSlope is the standard deviation of rising slopes, i.e. the slopes connecting a valley with the following peak. Rfilt means Relative Spectral Transform (RASTA)-style filtered. This means stddevRisingSlope of 5th Rfilt of auditory spectrum and it was smoothed using an averaging filter with window length 3.

5 Conclusion

Best results were obtained using the correlation based feature selection for both the full audience and reduced audience case. For the full audience experiment, the best results were obtained for Simon video with SVR and correlation based feature selection of 200 features. Again for reduced dataset experiment, the best results were obtained for Simon video with SVR and correlation based feature selection of 200 features. The SVR learning method consistently out-performed the MLP method.

The relevant audio features were also obtained for both the experiments. Future work can include further interpretation of the relevant features obtained. Also some other audio feature extraction methods can be explored.

References

[1] Zhang, John R., et al. *Correlating Speaker Gestures in Political Debates with Audience Engagement Measured via EEG*. Proceedings of the ACM International Conference on Multimedia. ACM, 2014.

[2] Eyben, Florian, Martin Wllmer, and Bjrn Schuller. *Opensmile: the munich versatile and fast open-source audio feature extractor*. Proceedings of the international conference on Multimedia. ACM, 2010.

[3] Schuller, Bjrn, et al. *The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism.* (2013).

[4] Eyben, Florian, et al. *The acoustics of eye contact: detecting visual attention from conversational audio cues.* Proceedings of the 6th workshop on Eye gaze in intelligent human machine interaction: gaze in multimodal interaction. ACM, 2013.

[5] Bellard, Fabrice, and M. Niedermayer. *FFmpeg.* http://ffmpeg.org (2012).

[6] Pedregosa et al. *Scikit-learn: Machine Learning in Python.* JMLR 12, pp. 2825-2830, 2011.

[7] M.W.Gardener, S.R. Dorling. *ARTIFICIAL Neural Networks(The Multilayer perceptron)- A review of the applications in Atmospheric Sciences.* June 1998.

[8] Christopher Bishop. *Pattern Recognition and Machine Learning.* book, (2007).

[9] Alex J. Samola, Bernard S. *A tutorial on Support Vector Regression.* NeuroCOLT2 Technical Report NCR2, October 1998.