COLUMBIA
UNIVERSITY

Project Report

COMS 6901

Correlation of Speaker Audio with Audience Engagement and Speakers Gesture

Author:
Deepak Ravishankar

Supervisor:
Dr. John Kender

Dec 31, 2018

Abstract

In this project we are trying to establish the relationship between the acoustic features of a speaker and amount of effect it has on the users engagement which is measured using the user’s score to the questions asked. For this purpose we build on the work done in the previous report and try to establish if there is some correlation between audio and the gestures. Multiple models were tested both from traditional machine learning and deep learning. Once a model of sufficient accuracy was found we used the data to establish if the voice modulation had any effect on the user and if it did have an effect, if it improved the scores or not.

Contents

1. Introduction

This projects tries to expand on the work done by Yang et al.[1] and tries to find further proof in the relation between acoustic features and the users engagement. The project used video recording of a lecture. For the video production, we invited a Columbia University professor to give the lectures and provided him with a set of PowerPoint slides with images and text along with a script that instructed him when to use specific gestures (metaphoric, iconic, deictic, beat) at particular points in the text. He was also told that he was free to use additional gestures where it felt appropriate. When the videos were produced, the lecturer had an audience and lectured quite naturally. His gestures included those we specified as well as those he added because they were part of his natural lecture style.

Further the videos were created into 3 different scenarios:

- With speaker, audio and slides
- With audio and slides
- With slides only

The slides were also in three topics of bicycles, perspectives and tarmacs. The test setup was designed in such a way so that the the hypothesis of , whether gestures affect a users retention of a topic or not, could be tested.

2. Data

For the experiments mainly the data from the bicycle scenario was used. Initially the videos were converted to audio data , i.e. from mp4 to wav using ffmpeg. wav was preferred due to it's lossless compression compared to other formats like mp3.

2.1 Preprocessing

Before the data could be used to train the models the audio was converted into clips of length 1 seconds which could be used by the VGGish model which could be used on the trained model.

3. Model

Instead of using defined features of the audio data like

- Energy
- Spectral
- Linear Predictive Coding (LPC)
- Perceptual Linear Prediction Cepstral Coefficients (PLPCC)
- Chroma feature vector
- Chroma deviations

We preferred to use derived features from a neural network as they may be better able to extract the signal from the noise in the audio data. To generate this derived features the VGGish network[3] was used. A pre-trained version of the model was used which had been trained on the Youtube 8M dataset[2]. The pre-trained model was taken from the tensorflow repositories. The output of the models penultimate layer were used as the derived features. The reason that a pre-trained network was used was due to the limitation of data. Also transfer learning has shown promise and thus features derived from a network trained on a different data set would be useful.

The network created a feature vector of length 128 in it’s second to last layer which were used as input to various machine learning models to be able to predict gesture audio pairs. A model trained on this would be able to just use the audio to predict the gesture and thus could be used to prove the correlation between audio and gesture and also highly correlated audio gesture pairs.

3.1 Feature selection

Due to the high dimensionality of the feature vector only the vector indices which showed a high correlation with the gestures were chosen. The 20 highest correlated features were chosen from the entire vector of length 128 and models were trained on both the 128 features and the 20 features. Though the models showed better prediction with 20 features due to the high dimensionality of the training data in case when all the features were used.

3.2 Model explorations

Various models were tried and both their test accuracies and confusion matrices were calculated. Below are the data on the models

Logistic regression

With all the 128 features the test accuracy was 0.45

Confusion matrix:

	Actual Gesture	Actual No gesture
Predicted Gesture	21	15
Predicted No Gesture	26	13

With all the 20 features the test accuracy was 0.58

Confusion matrix:

	Actual Gesture	Actual No gesture
Predicted Gesture	28	8
Predicted No Gesture	27	12

SVM with Linear kernel

With all the 128 features the test accuracy was 0.48

Confusion matrix:

	Actual Gesture	Actual No gesture
Predicted Gesture	36	39
Predicted No Gesture	0	0

With all the 20 features the test accuracy was 0.68

Confusion matrix:

	Actual Gesture	Actual No gesture
Predicted Gesture	26	10
Predicted No Gesture	18	21

SVM with RBF Kernel

With all the 128 features the test accuracy was 0.48

Confusion matrix:

	Actual Gesture	Actual No gesture
Predicted Gesture	24	18
Predicted No Gesture	21	12

With all the 20 features the test accuracy was 0.60

Confusion matrix:

	Actual Gesture	Actual No gesture
Predicted Gesture	26	8
Predicted No Gesture	30	11

K- Nearest Neighbors

With all the 128 features the test accuracy was 0.49

Confusion matrix:

	Actual Gesture	Actual No gesture
Predicted Gesture	32	9
Predicted No Gesture	26	7

With all the 20 features the test accuracy was 0.54

Confusion matrix:

	Actual Gesture	Actual No gesture
Predicted Gesture	34	7
Predicted No Gesture	31	2

3.3 Model selection

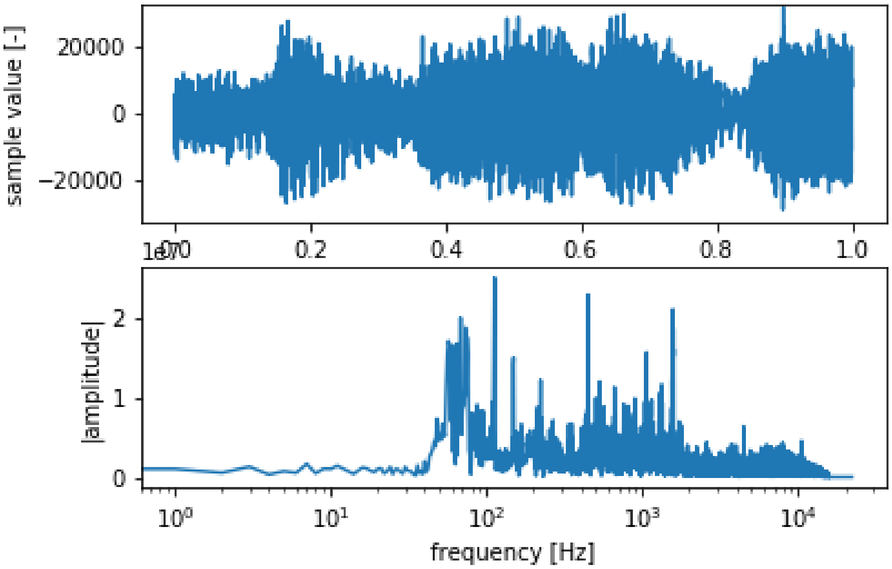
Comparing the above models we can see that SVM with linear kernel performs the best with a test accuracy of 0.68 which is high enough to establish that there is some correlation between the derived audio features and the users gesture. The confusion matrix is also spread to show that the predictions are not affected by the data being skewed.

4. Model explanation

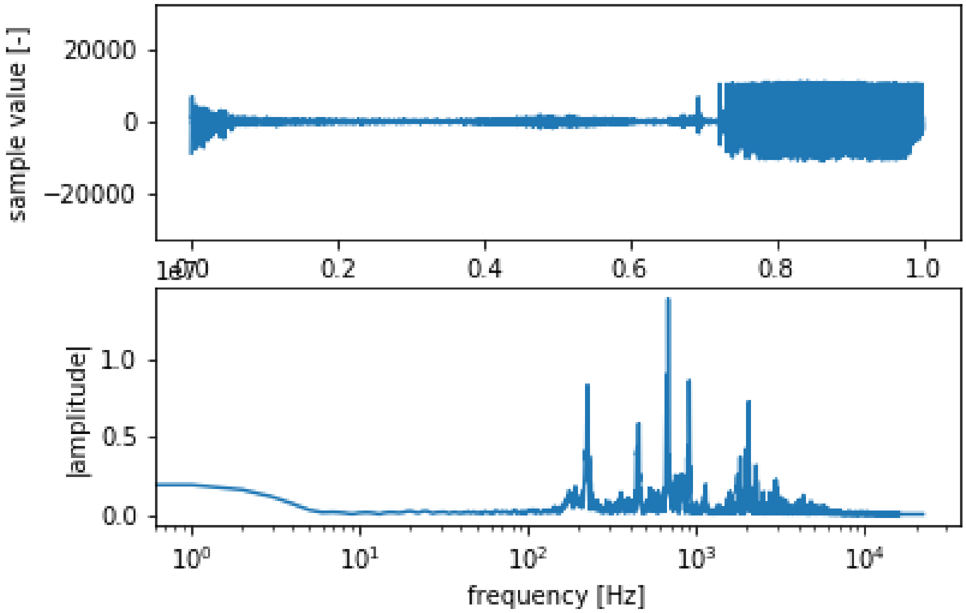
Now that we have established that there is some audio features which can be used to predict the gesture. We can try to explain what the model is trying to do. As there is a deep learning model in the model pipeline the findings in these section can only remain as Hypotheses.

To find audio clips which were similar to the audio clips which were successfully classified, we created the embeddings vector for the video in the Youtube 8m data sets and found the vectors with the lowest cosine distance to the successfully classified vectors.

These vectors were then manually and heuristically examined to find any similarities between the audio clips.



Frequency distribution of correctly classified vector



Frequency distribution of incorrectly classified vector

From the above figures and and the audio examinations we were able to establish the modulation of frequency and amplitude and important features which relate the audio and the speaker's gestures.

5. Correlation to scores

Looking at the question asked for the bicycle video for the candidates. We can see that only questions 1 to 6 have answers at particular timestamps and question 7 is a question based on the understanding of the video and hence cannot be used for correlation with audio

	Q1	Q2	Q3	Q4	Q5	Q6
mean	0.7704	0.639344	0.590164	0.704918	0.819672	0.377049
Std dev	0.4240	0.484176	0.495885	0.459865	0.387651	0.488669

Spread of scores of the questions

The model that was trained was correctly able to predict the gesture of questions 1,2 and 5. Whereas the gesture used on question 3,4 and 6 were not predicted and thus the audio in those questions is not as highly correlated

Looking at the scores we can see that the scores of questions 1,2 and 5 and overall higher than those of questions 3,4 and 6.

To verify this we perform t-tests on all the experiments and then we try to find the the validity of this hypothesis

Question pair	p-value
1,3	0.020
1,4	0.418
1,6	2.09×10^{-5}
2,3	0.616
2,4	0.398
2,6	0.006
5,3	0.003
5,4	0.127
5,6	7.29×10^{-8}

P-values for proving question in (1,2,5) is better than than (3,4,6)

From these we are able to see that the p values are significant for all question pairs except those involving question 2 and thus this can be hypothesis can be expected to a certain extent but it still needs to be made clearer by using using more examples

6. Conclusion

We were able to prove that there is a correlation between the audio features and speaker gestures using the derived features from the VGG model. After this we were able to use the strongly correlated models and show that the models which can be classified correctly seem to have some positive correlation with the test score and thus positively affect the retention of user to some extent

7. Future work

To further expand on this project we could try to make run the experiments on more examples. We could even try to make the experiments more substantial by improving the sample size. Further more robust statistical tests could be used to get clearer insights on the data. Models could be developed based on the various gesture types and then more insights could be drawn for the gesture types.

References

[1] Hang Yang,Correlation of loudness of speakers with Audience Engagement and speakers Gesture, Research Report, 2017
[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan: “YouTube-8M: A Large-Scale Video Classification Benchmark”, 2016
[3]Shawn Hershey et al.: “CNN architectures for large-scale audio classification”, 2017