

Unsupervised Speaker Cue Usage Detection in Public Speaking Videos

Anshul Gupta

Dinesh Babu Jayagopi

Multimodal Perception Lab
International Institute of Information
Technology, Bangalore
India

Abstract

A speaker's bodily cues such as walking, arm movement and head movement play a big role in establishing engagement with the audience. Low level, pose keypoint features to capture these cues have been used in prior studies to characterise engagement and ratings, but are typically not interpretable, and have not been subjected to analysis to understand their meaning. We thus apply a completely unsupervised approach on these low level features to obtain easily computable higher level features that represent low, medium, and high cue usage by the speaker. We apply our approach to classroom recorded lectures and the significantly more difficult dataset of TED videos, and are able to positively correlate our features to human interpretable ideas of a speaker's lateral head motion, and movement. We hope that the interpretable nature of these features can be used in future work to serve as a means of feedback to speakers, and to better understand the underlying structure behind the results.

1 Introduction

Public speaking has existed as a method of information dissemination since centuries. Good public speakers rely on a combination of both verbal and non-verbal cues in order to engage with their audience. Non-verbal cues play an important part in human interaction with a person's gait, arm movement, and head movement telling a lot about the speaker's emotions, confidence and personality [1]. Different speakers use these cues to varying degrees, which helps form our perception of them.

Several works have tried to find the importance of cues for affective expression. Sawada et al. (2003) [2] found velocity, acceleration, and finger motion range as important hand and arm movement features for distinguishing different affective states. Busso et al. (2007) [3] evaluated the effect of head movement on perception of affect, and found head motion corresponding to different affective states to be characterized by distinct motion activation, range, and velocity.

Studies have used these features to assess public speaking performance, or to predict engagement and ratings. Chen et al. (2014) [4] used a multimodal approach to assess public speaking performance. Their feature set included the use of the mean and standard deviation

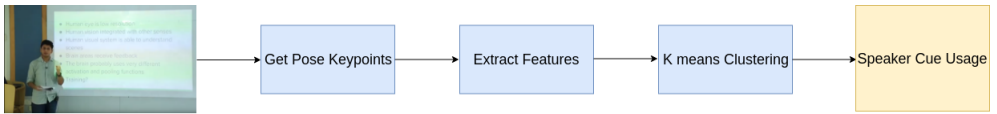


Figure 1: Approach Pipeline

of hip and hand movement speeds, and the mean and standard deviation of head movement. Cullen and Harte (2017) [6] calculated the speed and acceleration of hand movement of speakers, and used their mean and peak values as features to predict user ratings for thin slices of TED videos. However, these features are difficult to interpret and correlate to human interpretable ideas of walking, arm movement, and head movement. Further, the datasets used are typically videos consisting of classroom recordings and are not meant for public use. Videos of the latter variety tend to be more sophisticated and include cuts and zoom shots. In this work, we use a completely unsupervised approach to get a set of higher level features that quantize a speaker’s usage of cues. Specifically, we group speakers into three categories corresponding to low, medium and high usage of two cues: lateral head motion, and movement. Additionally, these features are cheap to compute and can be included as part of a study without much overhead. We compare our results with human annotations for the same for classroom recordings of lectures, and then apply our approach to the significantly more difficult dataset of TED videos. We highlight the important difference in the datasets and try our approach with modifications to account for the differences. Our results indicate positive identification of cue usage across both the datasets suggesting the possibility of using these higher level features in future work involving public speaking.

2 Approach

We first preprocess the data by clipping the videos to the last one minute. Previous studies [6] have shown that thin slices of video of upto one minute have high correlation with audience ratings. For the TED dataset we remove videos with the tags ‘Performance’ and ‘Live Music’, and those for which a person is detected for less than 10 seconds in the slice. We also remove outliers with extremely large feature values which would affect our results.

We use the OpenPose library [7] in order to get 2D pose keypoints from the speaker in the video. Openpose is a real time, multi-person keypoint detection and multi-threading library written in C++. It returns set of values corresponding to the X (horizontal) coordinate, Y (vertical) coordinate, and confidence score in detection of a person’s body parts.

We use these keypoints to get features that capture our cues. The features consist of the standard deviation, speed and acceleration of the keypoints. We use the mean value of the speed and acceleration across the slice in the case of the classroom recordings. In the case of the TED videos we first get the timestamps for different cuts in the slice using the PySceneDetect tool [8]. The standard deviation and mean values of speed and acceleration are then found for a cut, and the mean across cuts forms the final feature set. Each of the features are also normalized with respect to their max value. Finally, we cluster our data using K means clustering into 3 clusters and examine its contents to check how well they

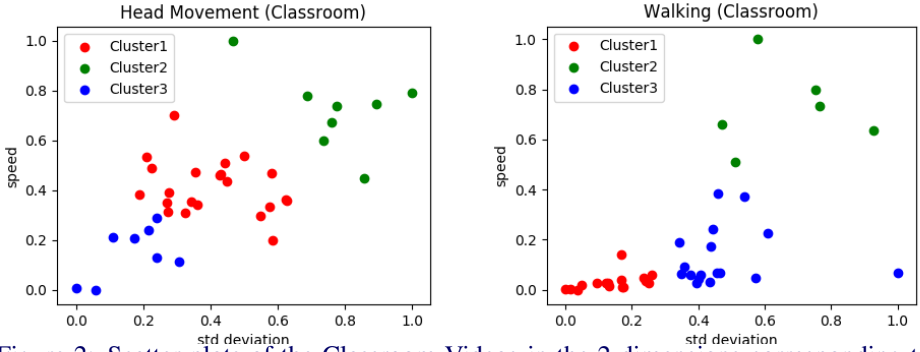


Figure 2: Scatter plots of the Classroom Videos in the 2 dimensions corresponding to the highest variance to represent the clusters (Best viewed in colour)

match with human annotations for speaker cue usage (Figure 1). The feature extraction algorithm for TED videos is explained in more detail in Algorithm 1.

2.1 Classroom Lectures

The classroom recorded videos consist of a set of 40 talks delivered by 29 different volunteers (15 female, 14 male). After removal of outliers, we had 38 videos for analysing head motion, and 39 videos for analysing movement. All the recordings were taken from a single fixed camera with no zooming or cuts. For most parts of the video, the participants entire upper body upto the knee is visible. We thus use the speaker’s hip X coordinate to characterize their movement and capture the cue of walking. To capture head movement, we use the speaker’s nose X coordinate, which is normalized by subtracting it’s value from their neck X coordinate. We get our feature set by using the approach highlighted in Section 3. To these datapoints, we apply K means clustering with 3 centers to get three distinct groups of videos (Figure 2).

2.2 TED Videos

TED videos are a set of public speaking videos aimed at information dissemination for the wider public on topics ranging from technology to entertainment. We downloaded these TED videos for a total of 1962 videos after removing videos with the tags of ‘live music’ and ‘performance’. After removal of outliers, we had 1946 videos for analysing head motion, and 1952 videos for analysing movement. Some of the key differences between TED videos and classroom recorded videos are:

- The camera following the speaker in close up shots
- Background of a single colour
- Multiple cuts and zoom shots in the video
- Occlusion of the body with often only the upper body above the shoulders visible
- Multiple people in the frame such as audience members in pan shots

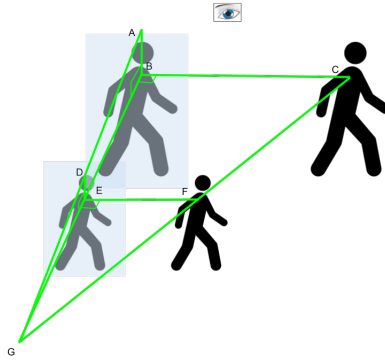


Figure 3: Proof behind the Inverse Zoom Factor

These differences make this dataset considerably more difficult to analyze. To circumvent these issues we use a number of techniques and heuristics.

As mentioned in Section 2, we first obtain a set of cuts using the PySceneDetect tool [8]. PySceneDetect is a command-line tool which analyzes a video, looking for scene changes or cuts. It returns the timestamp of each cut, after which we use the technique mentioned in Section 3 to obtain features. However, for these videos we scale each feature in a cut by an 'inverse zoom factor'. This inverse zoom factor is simply the inverse of the Y axis distance between the speaker's nose and neck. We choose these two anchor points due to the fact that these features will almost always be visible irrespective of the zoom as long as the speaker is in frame. We explain the logic for the above as follows:

In Figure 3, the larger shadows are closer, and the smaller shadows are further away from the camera. We want to normalize all movement at some distance from the camera to a corresponding amount of movement at a common distance from the camera; in other words for some movement EF at an arbitrary distance from the camera, we want to find the value of BC . From the figure it can be seen that the triangles GBC and GEF , and the triangles GAB and GDE are similar. This implies that:

$$\frac{BC}{EF} = \frac{GB}{GE}, \frac{GB}{GE} = \frac{AB}{DE} \quad (1)$$

Combining both we get that:

$$\begin{aligned} \frac{BC}{EF} &= \frac{AB}{DE} \\ BC &= \frac{AB}{DE} \cdot EF \\ BC &= \text{const.} \cdot \frac{EF}{DE} \end{aligned} \quad (2)$$

Also, since DE remains largely constant across a cut, we can extend our result for speed

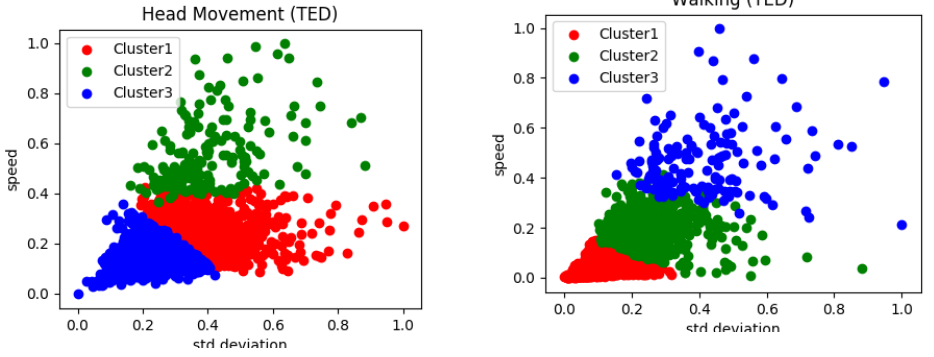


Figure 4: Scatter plots of the TED Videos in the 2 dimensions corresponding to the highest variance to represent the clusters (Best viewed in colour)

Algorithm 1: Feature Extraction (for TED)

Result: feature1, feature2, feature3

- 1 The pose keypoints, X ;
 - 2 The set of cut end timestamps, C ;
 - 3 The nose Y coordinates, $nose_y$;
 - 4 The neck Y coordinates, $neck_y$;
 - 5 $c_start = 0$
 - 6 **for** c_end in C **do**
 - 7 $nose_neck_diff = \text{mean}(nose_y[c_start:c_end] - neck_y[c_start:c_end])$
 - 8 $izoom[c_end] = \frac{1}{nose_neck_diff}$
 - 9 $c_start = c_end$
 - 10 **end**
 - 11 $c_start = 0$
 - 12 **for** c_end in C **do**
 - 13 $X_nonzero = \{x > 0 \mid x \in X[c_start:c_end]\}$
 - 14 **if** $\text{len}(X_nonzero) > 0.5 * (c_end - c_start)$ **then**
 - 15 $feature1 += \text{stddev}(X_nonzero) * izoom[c_end]$
 - 16 $feature2 += \text{mean_speed}(X_nonzero) * izoom[c_end]$
 - 17 $feature3 += \text{mean_acceleration}(X_nonzero) * izoom[c_end]$
 - 18 $c_start = c_end$
 - 19 **end**
 - 20 **end**
 - 21 $feature1 = feature1 / |C|$
 - 22 $feature2 = feature2 / |C|$
 - 23 $feature3 = feature3 / |C|$
 - 24 Normalize features by Max Value
-

and acceleration as well:

$$\dot{B}C = \text{const.} \frac{EF}{DE}, \ddot{B}C = \text{const.} \frac{\dot{E}F}{DE} \quad (3)$$

To account for the occlusion of the body, we use the speaker’s neck’s X coordinate to identify whether he has walked, and the speaker’s nose’s X coordinate subtracted from their neck’s X coordinate to identify lateral head movement. Further, we assume that the speaker will be the only person in frame for the majority of the video, and so take the keypoints of the first person observed. We apply Algorithm 1 on these keypoints and cluster them using K means clustering with 3 centers (Figure 4).

3 Results and Analysis

The distribution of datapoints across clusters indicates a clear separation into distinct classes (Table 1). Following intuition, we hypothesize that the cluster corresponding to higher values of features represents higher activity. Hence we sample points from the cluster closest to the origin, and furthest from the origin to capture high and low usage of the cue. We sample 5 points from each cluster in the case of the classroom videos and 25 videos from each cluster in the case of the TED videos, and position the problem as that of binary classification to our human annotators for high and low cue usage. Our accuracy scores are indicated in Table 2.

Cluster No. (in increasing distance from origin)	Head Movement (Classroom)	Walking (Classroom)	Head Movement (TED)	Walking (TED)
Cluster 1	8	16	1105	1232
Cluster 2	22	17	664	575
Cluster 3	8	6	177	145

Table 1: Distribution of datapoints across clusters

Data	Accuracy
Walking (Classroom)	60%
Head Movement (Classroom)	80%
Walking (TED)	58%
Head Movement (TED)	50%

Table 2: Accuracy scores across different cues and datasets

We observe high accuracy for identifying cue usage in classroom videos. The TED videos being more complex in nature show lower accuracy scores. Due to the high density of points in the cluster near the origin, increasing the number of clusters simply decreases the number of points in the cluster furthest away from the origin and we do not expect it to significantly affect accuracy.

Interestingly, despite better identification of head movement as compared to walking in classroom videos, our model performs worse for TED videos. This can also be seen in higher keypoint feature values for points in the cluster closest to the origin for head movement, as compared to the corresponding cluster for walking in Figure 4. On analysing the videos in

the high cue usage cluster, we notice that 16 of the 25 videos sampled had multiple people on the stage. Since Openpose returns the keypoints of all people detected, we notice that it switched the first person detected in a cut several times, whose coordinates we picked. This resulted in bloating the values of these videos, even if the main speaker does not move his head significantly. It also resulted in many videos with higher head movement being pushed to lower clusters. Head movement typically has lower keypoint feature values as compared to walking which is why this data was affected more.

An important point to note is that labelling cue usage is a subjective task, and our volunteers indicated lower confidence in their labelling of TED videos, further highlighting its complex nature. Despite the fact, our approach is able to capture usage of the cues in both datasets. Looking at cluster belonging, we can thus obtain a new feature that identifies the range of speaker cue usage into low, medium and high categories. The cluster distributions also give some insights into the general behaviour of speakers. As seen in Table 1, the number of speakers that incorporate a high degree of cue usage is small, with the majority belonging to the low usage category.

4 Conclusion

We have demonstrated a new approach to generating easily computable higher level features in public speaking videos that represent human interpretable ideas of walking and head movement. While results are still preliminary at this stage, future work that use these features to characterise engagement and ratings can benefit by providing useful feedback to the speakers regarding usage of cues. We also hope that by using these features, we can obtain insights into the results of studies that will help further improve our understanding for the same. Other directions of future work could involve grouping of speakers by style, where style can include high level interpretable visual features like the ones proposed.

References

- [1] Openpose. URL <https://github.com/CMU-Perceptual-Computing-Lab/openpose>.
- [2] Carlos Busso, Zhigang Deng, Michael Grimm, Ulrich Neumann, and Shrikanth Narayanan. Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1075–1086, 2007.
- [3] Brandon Castellano. Pyscenedetect, 2017. URL <https://github.com/Breakthrough/PySceneDetect>.
- [4] Lei Chen, Gary Feng, Jilliam Joe, Chee Wee Leong, Christopher Kitchen, and Chong Min Lee. Towards automated assessment of public speaking skills using multimodal cues. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 200–203. ACM, 2014.
- [5] Lei Chen, Chee Wee Leong, Gary Feng, Chong Min Lee, and Swapna Somasundaran. Utilizing multimodal cues to automatically evaluate public speaking performance. In

- 2015 *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 394–400. IEEE, 2015.
- [6] Ailbhe Cullen and Naomi Harte. Thin slicing to predict viewer impressions of ted talks. In *Proceedings of the 14th International Conference on Auditory-Visual Speech Processing*, 2017.
- [7] Daniel Gatica-Perez, L McCowan, Dong Zhang, and Samy Bengio. Detecting group interest-level in meetings. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 1, pages I–489. IEEE, 2005.
- [8] Fasih Haider, Fahim A Salim, Saturnino Luz, Carl Vogel, Owen Conlan, and Nick Campbell. Visual, laughter, applause and spoken expression features for predicting engagement within ted talks. *Feedback*, 10:20, 2017.
- [9] Michelle Karg, Ali-Akbar Samadani, Rob Gorbet, Kolja Kühnlenz, Jesse Hoey, and Dana Kulić. Body movements for affective expression: A survey of automatic recognition and generation. *IEEE Transactions on Affective Computing*, 4(4):341–359, 2013.
- [10] Mark L Knapp, Judith A Hall, and Terrence G Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.
- [11] Misako Sawada, Kazuhiro Suda, and Motonobu Ishii. Expression of emotions in dance: Relation between arm movement characteristics and emotion. *Perceptual and motor skills*, 97(3):697–708, 2003.