Automatic Identification of Non-Meaningful Body-Movements and What It Reveals About Humans

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

ABSTRACT

We present a framework to identify whether a public speaker's body movements are meaningful or non-meaningful (*Mannerisms*") in the context of their speeches. In a dataset of 81 public speaking videos from 27 individuals, we extract 5,027 instances of 350 unique body movement patterns (e.g. pacing, gesturing, shifting body weights, etc.). Online workers and the speakers themselves annotate the meaningfulness of the patterns. Five types of features were extracted from the audio-video recordings: disfluency, prosody, body movements, facial, and lexical. We use linear classifiers to predict the annotations with AUC up to 0.82. Analysis of the classifier weights reveals that it puts larger weights on the *lexical features* while predicting self-annotations. Contrastingly, it puts a larger weight on *prosody features* while predicting audience annotations. This analysis suggests public speakers tend to focus more on verbal features while evaluating self-performances. The audience, however, focuses more on non-verbal aspects of the speech.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; See http://acm.org/about/class/1998/ for the full list of ACM classifiers. This section is required.

Author Keywords

Behavioral Analysis; Public Speaking; Mannerisms; Data Analysis

INTRODUCTION

Public speakers often use body language to augment their verbal content. For example, hand gestures demonstrating "this big" may convey a sense of size. Gestures also help in communicating figurative information. For instance, when comparing two things, speakers tend to gesture as illustrated in Figure 1. This type of meaningful body movement reinforces the communication by adding to or modifying the content of a speech. Public speaking experts encourage these nonverbal behaviors [18, 41]. There are, however, other movements that do not add meaning to the speech, even though they appear frequently. Speakers may move inadvertently or gesture out of habit. Though often done unconsciously, such habitual movements may aid the articulation process [15, 17] (Imagine, for example, gesturing while talking in the phone. The other person does not see the gestures, but we still do it). In public speaking, these gestures can include self-touching, scratching, gripping, leaning, tapping fingers, rocking, swaying, pacing, playing with items in a pocket, adjusting hair or clothing, and more. Toastmasters International named this type of body movement "*mannerisms*" [40]. Mannerisms, as defined by Toastmasters, are often distracting to the audience, and most suggest avoiding them while delivering a speech [18, 41, 40].

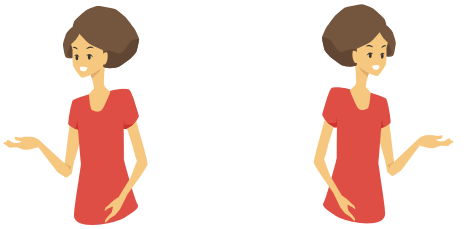

Figure 1. A comparison gesture

We present a computational framework to identify mannerisms from a multi-modal (Audio-visual and "MoCap" signals) record of public speaking. Automatic detection of mannerisms and showing those back to the speakers could be beneficial for the public speakers in becoming aware of their body language. Contemporary research shows that the conscious mind has a limited processing capability [23, 43] which is easily overwhelmed [11, 15] while speaking. Therefore, public speakers are typically unaware of their mannerisms while speaking. An automated mannerism detector could help people be more conscious of their body language without seeking out public speaking experts. Analyzing the characteristics of mannerisms could help further our understanding of fundamental aspects of human behavior.

There are several challenges in automatically detecting mannerisms. In order to recognize whether a specific gesture is meaningful within a speech, it is important to evaluate the context of the speech. For example, when a person gestures as in Figure 1, and compares two things—it is meaningful. But that same gesture may appear as mannerism in another context. It is currently difficult to detect this match because a) we do not have an exhaustive list that ties body movements to their corresponding verbal contexts, and b) it is difficult to

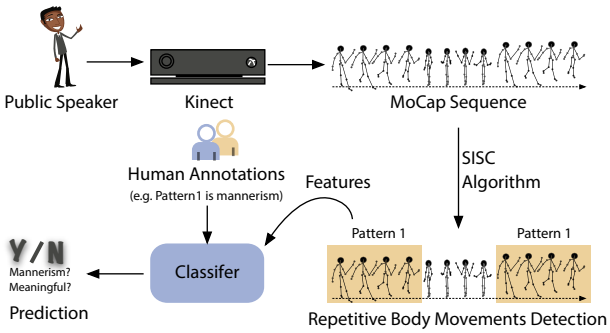

Figure 2. Detection of Mannerism in Public Speaking

accurately infer the verbal context from the uttered words. We address the challenges by utilizing our observation that the gestures annotated as mannerisms appear together with speech hesitations. Surveying the relevant literature reveals that mannerisms and speech hesitations stem from the same source. As a result, correlation is likely (More discussion in the "Related Literature" section.). We utilize this correlation characteristics by training our system to use features from both speech hesitations and body languages for detecting mannerisms.

The framework for automatic detection of mannerism is illustrated in Figure 2. We extract Motion Capture (MoCap) sequences for 81 public speeches by 27 speakers. A MoCap sequence records the full body movements of the speakers in three-dimensional coordinates. We use a *shift-invariant sparse coding* (SISC) algorithm [36, 37] to identify the commonly appearing *patterns* of body movements. By "pattern", we refer to a short (approximately two-second) segment of a MoCap sequence that appears frequently within the speech. Note that we are only interested in frequently-occurring body movements. A singly occurring movement is likely to be random. If a body movement appears several times, it can be either meaningful or mannerism depending on the context.

Three online workers assess each pattern, along with verbal context, to decide if the pattern is just a mannerism or is a meaningful gesture. The workers act as an audience member and evaluate the patterns accordingly. Similar annotations are collected from the public speakers immediately after they finish speaking. We also manually transcribed the speeches, including any filler words (uh, um, ah, etc.).

We extract five different types of features capturing speech hesitations or disfluency (e.g. use of filler words, prolonged silences, repetitions), facial expressions (looking directions, smile, head nod), prosody (pitch and loudness), body movements, and word use or lexical styles. We use several linear classifiers and regressions, such as LASSO, Max-Margin, and LDA (Linear Discriminant Analysis), to predict the human annotations from the extracted features. We also use them to interpret the relative importance of the features from the weight distributions. In addition, we use a neural network to model any non-linear dependencies among the features.

Our results show that it is possible to predict mannerisms with a much higher degree of accuracy than a random chance. The speaker's self-annotations, however, are comparatively more difficult to predict than the online workers' annotations. This suggests that these two annotation styles evaluate two different aspects of the speeches. In order to further understand the difference, we analyze the relative weight distributions in each classifier. The results show that the participants' self-annotations are highly predictive with verbal features. On the other hand, online workers' annotations are more predictive with non-verbal features. We interpret this result as a strong indication that speakers tend to focus more on *what* they are saying, but the audience focus more on *how* the speakers say it.

In summary, we detail the following key contributions:

- We propose a system to detect mannerisms by utilizing their co-occurrence with speech hesitations.
- We design the system to be able to predict mannerism from both the speakers' perspective and the audience's perspective.
- We quantify the differences between online workers' annotations and participants' self-annotations by evaluating the weight distributions of the trained classifiers.
- Our findings assert that speakers tend to focus more on what they say while the audience focuses more on how the speakers say it.

RELATED LITERATURE

In order to gain a better understanding of mannerisms, we surveyed the relevant theoretical aspects behind it. In this section, we discuss the literary backgrounds on mannerisms, its relationship with speech hesitations, and other works similar to this research.

Mannerisms and their Characteristics

Body language is effective when it conveys the same meaning as the verbal content it accompanies [18, 41]. When the gestures appear inconsistent with the verbal content, they are referred to as "mannerisms" [40]. Traditionally, gestures in speech communications are classified into four different categories: *iconic*, *metaphoric*, *deictic*, and *beat gestures* [21]. Iconic gestures are hand movements illustrating object attributes or actions (for example, demonstrating "big" by holding hands apart). Metaphoric gestures put abstract concepts into a more concrete form (forming hands into a heart shape represents love and affection, for instance). Deictic (or pointing) gestures are typically used to point to the location of an object. Beat gestures reflect the rhythms of speech and are typically used to help listeners direct the focus of attention to important information [3]. Mannerisms, however, don't fit these categories in the sense that they are not meaningful to the content of the speech. It does not convey any specific semantic information. Mannerisms are typically expressed inadvertently to cope with cognitively demanding situations [15, 17].

Mannerisms are distracting to audiences [18, 41]. Dick et al. [13] described an important phenomenon to partially explain why this happens. They used functional magnetic resonance imaging (fMRI) to examine the influence of gestures

on neural activity in brain regions associated with processing semantic information. Their experiment shows that the human brain is hardwired to look for semantic information in the hand movements that accompany speech. In other words, listeners use a significant amount of neural activity to decipher body movements in association with the verbal content. Mannerisms cognitively overburden the audience without conveying any semantic information. This might be a reason why listeners find mannerism distracting.

Production of Speech Hesitations and Mannerisms

In our manual analysis, we noticed that mannerisms appear with speech hesitations such as filler words and long pauses. In order to investigate the cause of this co-occurrence, we studied existing literature on the production of speech hesitations and mannerisms.

Plenty of evidences suggest that speech hesitations occur when the speakers are uncertain about what to say, or choose what to say (as a result of many choices) [11]. For instance, Sharon Oviatt [27] observed that disfluencies occur more often before longer utterances. The author showed that a simple linear regression over utterance length can account for 77 percent of variations in speech disfluency. Merlo et al. [22] showed that disfluencies occur more while talking about unfamiliar topics. Beattie et al. [2] observed that utterances with words described as *contextually improbable* are more likely to be disfluent. They suggested that speech disfluencies arise from an element of choice in selecting an appropriate word with low contextual probability. These studies strongly imply that disfluencies are likely to occur when the speakers are burdened with thinking, planning, or choice.

Stanley Grand [15] analyzed the characteristics and causes of a specific type of mannerism: *self-touching*. He concluded that these hand movements are a form of self-feedback that helps speakers reduce cognitive overload by narrowing their attention. Such feedback helps speakers articulate simple chunks of information. Jinni Harrigan [17] performed a more recent study during medical interviews with 28 physicians and their patients. This study provided strong evidence that self-touching is related to information processing and production.

It is clear from the experiments described above that speech hesitation is a phenomenon observed during speaking situations with a high cognitive load. In addition, there is evidence that some types of mannerisms are exhibited in order to cope with high cognitive loads. As both speech hesitation and mannerisms arise in similar situations, it follows that both might appear together during a speech. We use this intuition to predict mannerisms by designing features related to speech hesitations (disfluency and prosody, for example).

Similar Works

Much research has been conducted in order to build systems to help with various aspects of public speaking. Damian et al. proposed a system named "Logue" [12] to increase public speakers' awareness of their own nonverbal behavior. It utilizes various sensors to analyze the speakers' speech rates, energy, and openness. Results are communicated to speakers using a head-mounted display. Roghayeh Barmaki [1] proposed an online gesture recognition application that would provide feedback through different channels, including visual and haptic channels. Bubel et al. created "AwareMe" [6] to provide feedback on pitch, use of filler words, and words per minute. AwareMe uses a detachable wristband to provide feedback to speakers while practicing. A Google Glass application named, "Rhema" [35] can provide real-time feedback to public speakers about their prosody (speech rate and volume). It was designed to reduce distractions to the speakers by providing succinct visual feedback. Similar efforts have been made by Luyten et al. [20], who explored the possibility of designing feedback systems by putting displays close to users' peripheral fields of vision. Such systems communicate real-time information to users effectively.

There are several kinds of research on detecting competence in public speaking. Chen et al. [9] proposed a multimodal sensing platform for scoring presentation skills. They used syntactic, speech, and visual features (e.g. hand movements, head orientations, etc.) with supervised regression techniques (support vector regression and random forest) in order to predict a continuous score for public speaking performance. They claimed a correlation coefficient of 0.38 to 0.48 with the manually annotated ground truth.

Only a few research focus on the body language in public speaking. Nguyen et al. [26] implemented an online system to provide feedback on a speaker's body language. The feedback is given on a five-point scale. The authors recorded physical movements using the Kinect, then used a nearest-neighbor classifier to compare the recorded movements with a set of predefined templates of body movements. The templates contain ground truth information about the possible feedback. As it uses nearest-neighbor classification, this method cannot provide appropriate feedback for novel body movements. Additionally, it is not applicable for assessing the contextual relationship of gestures to verbal content.

"AutoManner" [37] is an interactive system to make public speakers aware of mannerisms. The authors extracted the speakers' repetitive body movements (patterns) and designed an interface to show these patterns to the speakers. The results show that question-and-answer-based interaction can make speakers aware of their mannerisms. The system could not, however, automatically detect which of the repetitive patterns were mannerisms. Without any automated prediction mechanism, the speakers need to go through many instances of the repetitive patterns in order to identify potential mannerisms. In this paper, we implement a computational framework to detect mannerisms automatically. This framework turns out to be an interesting tool for gaining insights into human behavior.

DATASET

In order to study mannerisms, we collected a public speaking dataset. It consisted of speeches from 27 speakers—14 females and 13 males. All speakers were undergraduate or graduate students and native English speakers. Each speaker spoke three times, approximately three minutes each time. The speakers were allowed to choose their own topics while providing a general guideline that the topics should be easily

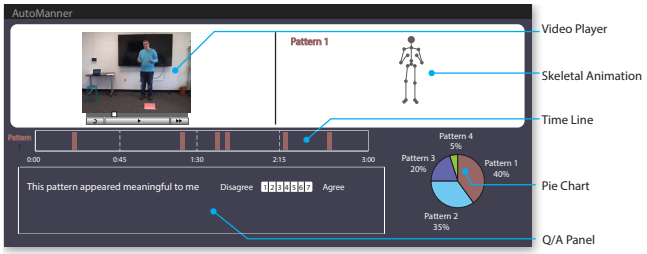

Figure 3. Screenshot of the ground truth annotation interface

understandable to general audience. The most common topics were—favorite book, movie, computer-games, hobby, superhero, personal idol, passion etc. Speech topics were decided two days earlier to allow them time for preparation. During the speech, the speakers' full body movements were captured using a Kinect depth sensor. Microsoft Kinect SDK [32] was used to extract the three-dimensional coordinates of 20 body-joint locations. We call the sequence of these joint coordinates a *MoCap* sequence. We used an SISC [35] algorithm to extract common body movement patterns from the speeches. We could extract 350 unique patterns of body movements using the algorithm; which appeared a total of 5,027 times in various speech videos of the dataset. This algorithm is described in the "Technical Details" section. The public speeches were recorded with a high-definition video camera. We manually transcribed the speeches, including all filler words (uh, um, ah, etc.).

Annotations

The participants watched their own speech during the analysis phase. Then, they used the interface shown in Figure 3 to annotate mannerisms. The interface contains several components: a skeletal animation, a timeline, a video player, a pie chart, and a question-and-answer box. The skeletal animation shows a pattern of body movements extracted by the SISC algorithm. The timeline highlights the instances of the pattern. If an instance is clicked on the timeline, the video player plays a clip of several seconds from the video centering around the clicked instance. This gives the participants an opportunity to recognize the context of the speech while analyzing the body movement patterns. Speakers rated on how meaningful the body movement patterns were in the context of the speech. The ratings were collected on a seven-point Likert scale, where seven indicates "very meaningful". Before collecting the annotations, the users were introduced about the interface and the analysis process using a recorded demonstration video.

Each speaker only analyzed and annotated mannerisms in one of their three public speaking videos. To encourage for a large amount of participation, participants only had to come to the lab once to complete all three speeches. However, a full session already took approximately 45 minutes to complete. It was not feasible to run the SISC algorithm over all three videos where we can occupy the participants for a longer period of time.

The same interface was used to gather annotations from the online workers. The workers—called "Turkers"—were recruited on the Amazon Mechanical Turk website. We selected the

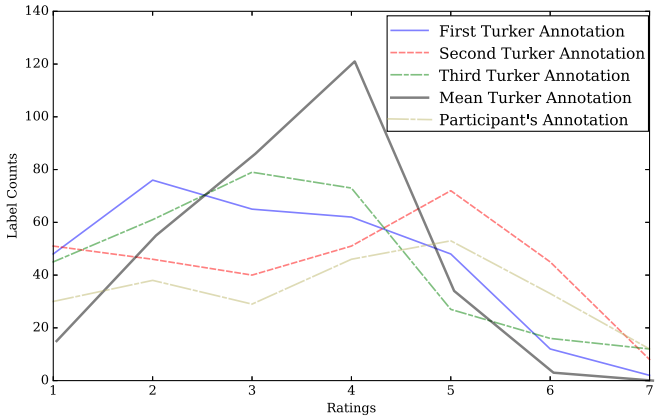

Figure 4. Distribution of labels for various annotators

Turkers located in united states only. In order to assure the quality of annotations, we selected only the Turkers having an experience of completing at least 500 jobs and 95 percent acceptance rate. The Turkers annotated all the videos in the dataset, including ones skipped by the participants. For reliability issues, three different Turkers annotated the same video. In total, 30 different Turkers participated in the annotation task.

Distribution of Annotations

Figure 4 shows the distribution of the annotated labels over a range of one to seven. The distributions are shown for individual Turkers' annotations, participants' self-annotation, and the mean of the Turkers' annotations. The mean distribution follows a normal-like shape due to the central limit theorem. Note that the total number of participants' annotations (the area under the participant distribution curve) is much smaller than the total number of Mechanical Turk annotations. This is due to the fact that the participants annotated only one of their three speech videos.

TECHNICAL DETAILS

We design a computational framework in order to answer the following questions:

1. Is it possible to automatically detect mannerisms? If so, to what extent?

2. Is there any difference between self-judgment and audience judgment in the annotations of mannerisms? If so, what is it?

It is possible to manually observe the videos in the dataset to formulate a hypothetical answer to these questions. However, manual analysis is costly in terms of time, effort, and money. It is also difficult to reproduce such experiments. Therefore, we perform a computational analysis of mannerisms. With such a technique, it will be possible in the future to apply the same analysis on a new dataset—validating our results in new contexts. In this section, we discuss various components of the computational framework as outlined in Figure 2.

Algorithm 1: Extracting the repetitive patterns

Input: $\mathbf{f}[n]$, M, D and λ
Output: ψ, α
Initialize;
$\alpha \leftarrow 0$, $\psi \leftarrow$ random;
while *notConverge* **do**
  Update ψ using Gradient Descent;
  Project ψ into the feasible set, $\{\psi : \|\psi\|_{\mathrm{F}}^2 \leq 1\}$;
  Update α using Gradient Descent;
  Shrink α to enforce sparsity;
  Project α to feasible set, $\{\alpha : \forall_n \alpha[n] \geq 0\}$;

Extracting Repetitive Patterns

We record the Motion Capture (MoCap) sequence of the speakers' body movements using a Kinect sensor. We extract the patterns of body movements using the SISC algorithm[1] as implemented by Tanveer et al. [36, 37]. It is an unsupervised algorithm for extracting frequently occurring contiguous segments from a MoCap sequence. Being unsupervised, the algorithm does not require a list of possible templates for extracting the patterns. It works by solving the optimization problem shown in equation (1).

$$\begin{aligned} \hat{\psi}[m], \hat{\alpha}[n] = & \arg\min_{\psi,\alpha} \frac{1}{2}\|f[n] - f_{\text{model}}[n]\|^2 + \lambda\|\alpha\|_1 \\ & \text{s.t.} \quad \|\psi\|_F^2 \leq 1 \quad \text{and,} \quad \forall_n \alpha[n] \geq 0. \end{aligned} \tag{1}$$

In this equation, $f[n]$ represents the MoCap signal captured from the Kinect sensor. $f_{\text{model}}[n]$ represents a mathematical model of the MoCap signal, as shown in Equation (2).

$$f_{\text{model}}[n] = \sum_{d=0}^{D-1} \alpha_d[n] * \psi_d[m] \tag{2}$$

Here, ψ_d represents one among a total of D possible patterns. α_d represents the corresponding locations where ψ_d appeared. α_d can be thought of as a train of impulse functions. The asterisk sign represents the convolution operation. The solution approach is shown in Algorithm 1. For further detail, please refer to the works of Tanveer et al. [36, 37].

Feature Extraction

We enlist the features we considered for classification and regression tasks in Table 1. We extract all these features from the region within the time-frame where a specific pattern appeared.

Disfluency Features

We discussed earlier that mannerisms tend to appear together with speech hesitations. Therefore, in order to detect mannerisms, we design features capturing the speech hesitations. We compute the average time a speaker talks, uses filler words, and remains silent, all while showing a specific pattern of body movements.

We compute these features by aligning transcripts of the speeches with the corresponding audio clips using the Penn Phonetics Lab forced aligner [44]. The transcript was human-generated and includes all filler words. After alignment, we compute the average time the speakers take to say words, to remain silent, and to utter filler words while showing a pattern of body movement. We also compute the counts of unique words, filler words, pauses, and their relative proportions. These features capture many important characteristics such as speaking rate, repetitions etc.

[1] SISC code is available in https://github.com/ROC-HCI/AutoManner

Table 1. Features extracted

Category	Feature Names	Count
Disfluency	Mean of speaking time, filler word time, and pause length; Count of unique words, filler words, pauses; Relative proportions of words, fillers and pauses	9
Prosody	Avg.,Min,Max,Range and Standard deviation of loudness, pitch, and 1st, 2nd, and 3rd formants; Ratios of voiced to unvoiced regions	26
Body Movements	Mean and std of position, velocity, and acceleration of the elbows, wrists, knees, and ankles	40
Face	Pitch, Yaw, and Roll of head; Normalized distances between various points of face	24
Lexical	Counts of words in 23 LIWC categories	23

Prosody Features

Prosody features contain the intonation patterns, loudness, and formants of the vocal sound. It captures the speaking style of the speaker. In addition, it can be a good indication of the speaker's affective states; for instance, boredom, excitement, or confusion. We include prosody in our computational framework as confusion is a dominant cause of speech hesitation. Prosody has been found useful in many relevant research—intent modeling [34], job interview performance prediction [25], etc. We use PRAAT [5] to extract the loudness, pitch, and the first three formants from the speech signal. Then, we compute summary statistics (mean, standard deviation, etc.) of these prosody signals from the two-second segments of each pattern.

Body Movement Features

We wanted to extract some features related to pace and movement, as mannerisms are related to body language. These features are extracted from the MoCap sequence captured using the Kinect. We calculate the position, velocity, and acceleration of the speakers' elbows, wrists, knees, and ankles with respect to a reference point on their body. We calculate the mean and standard deviation of the velocity and acceleration. We also calculate the mean positions of the joint locations. These features capture the amount of movements in each pattern.

Facial Features

Face is, perhaps, the most important channel for non-verbal communications. Facial expressions can encode many different affective states, including confusion and stress. As

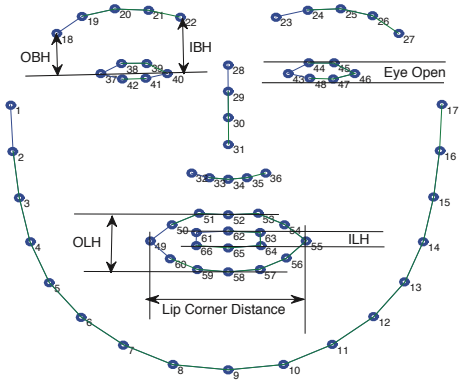

Figure 5. Illustration of facial features: OBH (outer eye - brow height), IBH (inner eye - brow height), OLH (outer lip height), ILH (inner lip height), eye-opening, and LipCDT (lip corner distance).

mannerism stems from high cognitive load situations, facial expressions could be relevant for detecting mannerisms. We extract low-level facial movements to capture any possible information relevant to mannerisms. Facial features are found useful in other research as well [29, 24].

We use a facial point tracker, proposed by Saragih et al. [31, 30], to track 66 landmark points on the face. From these points, we calculate the pixel distances of OBH, IBH, OLH, ILH, eye-opening, and LipCDT, as illustrated in Figure 5. These distances are normalized by the distance between a subject's pupils in order to remove any scaling (zooming) effect. We compute the mean and standard deviation of these distances within the two-second length of each pattern. In addition, we compute the mean and standard deviation of the pitch, yaw, and roll movements of the head using the tracker. These features can capture where a person is looking. In our manual observation, we found looking away from the audience while showing repetitive gestures is a good feature for detecting mannerisms.

Lexical Features

In order to capture information about the verbal content of the speech, we extract some lexical features. We use a psycholinguistics tool named *Linguistic Inquiry Word Count* (LIWC). LIWC describes 64 different categories of positive and negative emotions (happy, sad, angry, etc.), function word categories (articles, quantifiers, pronouns, etc.), content categories (anxiety, insight, etc.), and more. We use *greedy backward elimination feature selection* [7] to choose the 23 most relevant categories as lexical features.

Feature Normalization

The features we select have different dynamic ranges. To make sure that one group of features does not dominate others, we apply z-score normalization over the features. In other words, we subtract the mean of a specific feature from each feature value and divide by the standard deviation. This makes the features be distributed with zero mean and unit variance.

Classification and Regression Analysis

We use four different types of classification and regression techniques to compare their relative performances in predicting mannerisms. We use three linear classifiers (LASSO, LDA, and max-margin) in order to compare the relative proportions of feature weights. The feature weights can provide us valuable insights on what makes a repetitive movement meaningful or, conversely, a mannerism. We use a nonlinear classification and regression method (a neural network) to see if there is any improvement from a linear classifier.

LASSO

The *least absolute shrinkage and selection operator* (LASSO) [39] is a regression technique that jointly performs variable selection and regression. As we selected a large number of features from various channels of verbal and non-verbal communication, it is likely that some are correlated. LASSO automatically chooses the best feature and suppresses correlated features. This improves the regression performance, as it is less likely to be affected by the *the curse of dimensionality*. We use the LASSO implementation of scikit-learn [28] for regression. For LASSO classification, we put logistic (sigmoid) function over a linear predictor with $\ell 1$ regularization. We use Keras [10] and Theano [38] to solve the corresponding optimization problem.

Max-Margin

The max-margin classifier, or *support vector machine* (SVM) [42], is a popular classification technique that maximizes the margin between two classes. It does not, however, perform automatic feature selection. We use the linear SVM implementation from the LibSVM library [8]. Additionally, we use the support vector regression (SVR) [33] from the same library for maximum margin regression.

LDA

Linear discriminant analysis (LDA) [14] projects the data on to a maximally separating hyperplane for classification. A maximally separating hyperplane could emphasize the differences between meaningful and non-meaningful gestures in terms of feature distributions. However, unlike LASSO, it does not perform feature selection. We use scikit-learn [28] for both LDA classification and regression.

Neural Network

Some features may depend on each other to represent mannerisms. Such interdependency is difficult to model with linear classifiers. We use a feed-forward neural network in order to capture a possible non-linear decision surface. In our network, we apply two hidden layers, each containing 16 nodes. The output layer contains only one node. We use a *rectified linear unit* (ReLU) activation for the hidden layer. The output layer contains either a sigmoid or a ReLU activation, depending on whether it is a classification or regression task, respectively. We use Keras [10] and Theano [38] for implementation of the network.

RESULTS

In this section we discuss the classification and regression performances. In addition, We discuss some experiments we conducted for better understanding of the results.

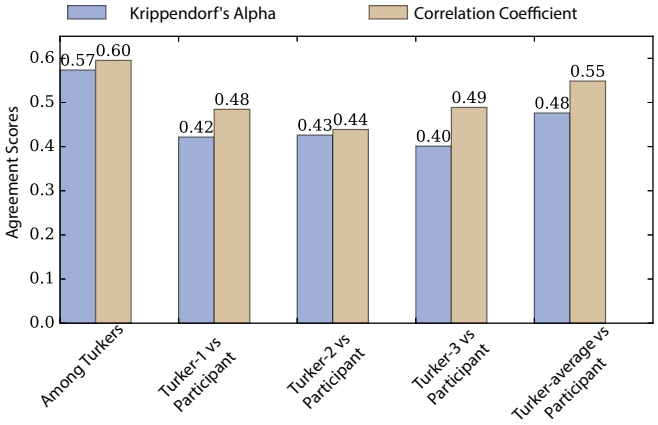

Figure 6. Inter-Rater agreements among various annotators.

Inter-Rater Agreement

Figure 6 shows the inter-rater agreement among the annotators. The agreement is calculated using Krippendorff's alpha [19] and correlation coefficient. Krippendorff's alpha allows calculation of agreement among any number of raters, even with missing data and various types of measurements (binary, nominal, ordinal, interval, and more). In our experiment, we collect the annotation on a seven-point scale, which is an interval datum. In addition, annotations are collected from more than two annotators. Therefore, Krippendorff's alpha is the ideal choice for measuring agreements in our experiment. The figure shows that there is moderate ($\alpha = 0.57$) agreement among the Mechanical Turk annotators. The individual Turkers, however, do not agree as much with the participants' annotations as they do among themselves. This provides the first indication that the Turkers' opinions were different than the participants' self-reflections.

Classification Performances

Table 2 shows the performances of various classifiers in distinguishing mannerisms from meaningful gestures. It represents the area under the ROC curve (AUC) [16] as a measure of classification performance. The AUCs are averaged across 30 measurements on randomly-split training and test subsets. As there are only two classes in this task, a completely random classifier would have shown an AUC of 0.5. LASSO performs better than random classification, as it shows an AUC of 0.82. The "MTurk Annotation" column represents the AUCs when we use all the data points from the Mechanical Turk dataset. We also randomly sub-sampled the Mechanical Turk dataset into one-third of its original size in order to match the size of the self-annotated dataset. AUCs of the classifiers on this sub-sampled dataset are shown in the "Subsampled MTurk Annotation" column. Finally, the classifier performances on the self-annotation dataset are shown in the "Self Annotation" column.

It is evident from Table 2 that, for all the classifiers, the Turkers' annotations are easier to predict than the participants' self-annotations. It happens consistently for all classifiers, even when the sub-sampled MTurk dataset is used. We performed statistical t-tests between the prediction results of self-annotations and subsampled MTurk annotations. The

Table 2. Area Under the ROC Curve (AUC) for various classifiers

Classifier	MTurk Annotation	Subsampled MTurk Annotation	Self Annotation
LASSO	**0.82**	**0.69**	**0.65**
Max-Margin	0.78	**0.69**	0.60
LDA	0.77	0.73	0.63
Neural Network	0.71	0.60	0.59

Table 3. Correlation Coefficient for various regression methods

Regressor	MTurk Annotation	Subsampled MTurk Annotation	Self Annotation
LASSO	**0.63**	**0.55**	**0.37**
LDA	0.59	0.38	0.31
Neural Network	0.48	0.35	0.28
Max-Margin	0.35	0.20	0.05

difference between these two groups is statistically significant ($p << 0.01$) in all the cases. Therefore, the results indicate, the difference between self-annotations and Turkers annotations is not just an anomaly of classification procedure. It is not even an effect of larger Turker dataset. This suggests that the participants' annotations are qualitatively different than the Turkers' annotations. We discuss this effect more in the "Discussions" section.

Regression Performances

Regression analysis is a stricter prediction approach than classification. There are infinitely many possible outcomes in regression, as it generates outputs in real numbers. In order to measure the performance, we calculate the correlation coefficient between the regression output and the human annotations. Table 3 shows the performance of the regression methods in predicting mannerisms. The annotations are collected on a seven-point Likert scale, where seven indicates highly meaningful and one indicates not meaningful at all. In this experiment, the maximum correlation coefficient we obtained is 0.63. The regression results show a similar trend as in Table 2: The Mechanical Turk annotations are more predictable than self-annotations. LASSO still performs best in the regression task, and max margin's performance is particularly poor. More specifically, while predicting the self-annotations, the max-margin regression performance is nearly equivalent to a random predictor. We discuss this more in the "Discussion" section.

Distribution of weights

The results so far strongly indicate that the speakers' self-reflections are different than the audience's (the Turkers') opinions. In order to understand the qualitative difference between these two, we look into the feature weights assigned by the linear predictors (classifiers and regressors). Analyzing the weights helps us identify how various features contribute to the prediction task. We calculate the weight of each category (prosody, disfluency, etc.) using equation (3),

$$W_c = \frac{1}{N_c} \sum_{f \in c} |W(f)| \tag{3}$$

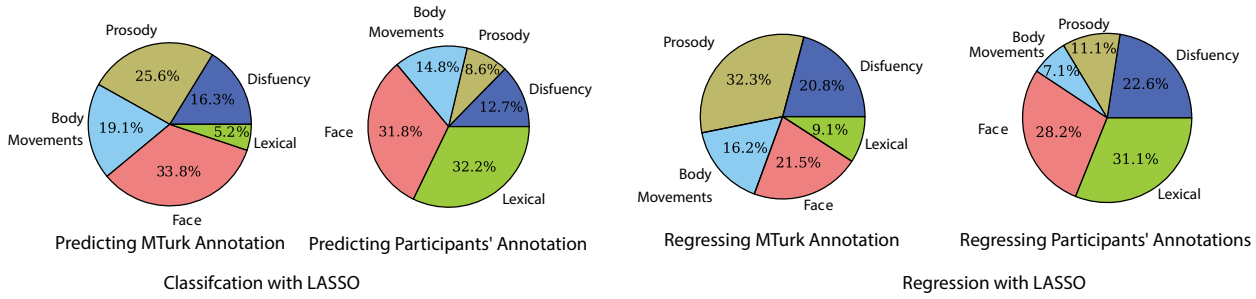

Figure 7. Distribution of various types of feature weights in the trained classifier

where c is a feature category, N_c is the number of features with nonzero weights in c, and $W(f)$ is the weight of the feature f. The names of the features are given in Table 1.

In Figure 7, we show the relative proportions of these normalized weights in pie charts. Notice that the lexical features take only 5.2 percent of the total weights to classify the Turkers' annotations but 32.2 percent to classify the participants' annotations. A similar trend is visible in the regression task, as well. Lexical features contribute only 9.1 percent of the total weights for Turkers' annotations. The same (lexical) features take 31.1 percent of the weights while regressing the speakers' self-annotations. In other words, lexical features are more predictive of the participants' annotations than they are for the Turkers' annotations. Notice, however, that the prosody and body movement features show an opposing trend. They contribute more while predicting the Turkers' annotations, but less while predicting the speakers' annotations.

To be sure this is not just coincidence, we observe the weights of the other two linear classifiers, max-margin and LDA, as well. Table 4 shows the distributions of the feature weights for various cases. "MTF" and "MTS" represent prediction of all Mechanical Turk annotations and their random subsamples. "Self" represents prediction of the speakers' self annotations. Note that, in most cases, lexical features take a higher percentage for self annotations than the Turkers' annotations. This is consistent with earlier observations. The only exception is when max-margin is used for regression purposes. We discussed previously, however, that max-margin was a poor regressor for predicting self annotations. Assigning comparatively lower weights on the lexical feature might be responsible for the poor performance.

This trend in feature-weight distribution suggests an interesting phenomenon. We see the data strongly suggests that participants are more observant about the verbal contents of their speech, whereas online workers notice the speakers' nonverbal body language more. This hypothesis seems plausible if we consider the amount of effort the speakers exert to produce the verbal content.

Notice that another prominent characteristic of the weight distribution is that the facial and disfluency features are almost consistently good predictors of mannerisms. We discussed disfluency (speech hesitations) earlier. It is also natural for facial expressions to be good predictors as it is one of the major nonverbal cues.

DISCUSSIONS

The results of our experiments provide at least three evidences that show that annotations from the public speakers and the online workers are different. **Firstly**, the inter-rater agreements show that the online workers agree more among themselves than they do with participants. **Secondly**, the performances of the predictors reveal that the online worker's annotations are much easier to predict than the self-annotations. **Thirdly**, the distribution of weights shows that there is a clear difference in how the features contribute to predicting the two different kinds of annotations. The speakers put more emphasis on the verbal features for determining if a gesture is a mannerism. The audience, on the other hand, emphasizes non-verbal aspects (prosody and body movements, among others) for detecting mannerisms. This might be another reason why speakers are typically unaware of their own body language.

The distribution of feature weights provides a clue on to why the classification and regression techniques performed poorly in predicting self-annotations. The weights imply that the participants focus less on the non-verbal aspects of the speech—they put more emphasis on the verbal aspects. We did not, however, extract many verbal features in this experiment. The LIWC features capture only a statistical distribution of the words in a "Bag of Words" model. It does not capture any syntactic information. Many important mannerism cues might be captured in grammatical accuracy, sentence formulation, or even in the stylistic aspects of the speech. It may be possible that the features that we selected did not include the full spectrum of verbal qualities. However, it is also possible to capture additional nonverbal features including eye contact, high level interpretable facial expressions (e.g., surprise, happy, concentration, thinking etc.) and characteristics of pauses (e.g., using pauses to appropriately building up a suspense). Our future work will involve adding more verbal and nonverbal features to further validate the findings of predicting mannerisms.

Although LASSO and max-margin performed comparably well in the classification task, LASSO outperforms max-margin for regression. (LASSO performed better than other techniques, too, in both classification and regression tasks.) This is expected, as LASSO is naturally a regression technique and max-margin is a classification technique. In addition, there is a sparsity constraint in the formulation of LASSO. Due to this constraint, it gains the benefit of feature selection when several features are highly correlated. Selecting a limited number of features helps fight the "curse of dimensionality" [4]

Table 4. Percentages of Weights in Various Classification and Regression Techniques

Method	Task	Annot.	Percentages of Weights for various Feature Categories				
			Disfluency(%)	Prosody(%)	Body(%)	Face(%)	Lexical(%)
Max-Margin	Classification	MTF	10.4	26.3	17.7	26.4	19.3
		MTS	11.9	39.9	16.1	15.5	16.6
		Self	18.3	13.0	7.4	26.9	34.5
	Regression	MTF	20.7	19.1	16.8	21.1	22.3
		MTS	27.8	16.8	13.9	18.6	22.8
		Self	21.4	18.8	16.8	22.0	20.9
LDA	Classification	MTF	19.0	20.5	17.5	22.6	20.3
		MTS	20.1	21.9	14.6	21.6	21.7
		Self	20.9	17.0	14.1	23.2	24.8
	Regression	MTF	21.5	19.6	19.0	19.5	20.4
		MTS	22.3	19.2	17.0	18.8	22.7
		Self	20.9	16.8	15.5	21.4	25.4

problem. As we chose a lot of features, it is likely that some are highly correlated. This makes LASSO an apt approach for predicting mannerisms.

CONCLUSION

In this work, we proposed a computational approach to automatically detect mannerisms when a public speaker presents a speech. This method can detect mannerism with a reasonable accuracy (AUC up to 0.82). In addition, the proposed system can detect mannerisms from both the speakers' and the audience's perspectives. This system can be useful in making public speakers aware of their body language.

In addition, deeper analysis of the prediction methods provided interesting insight into human behavior. Our results indicate that the way a speaker evaluates his/her own speech is different than the way an audience perceives it. Speakers tend to emphasize more on verbal aspects of the speech, while the audience focuses on the non-verbal aspects. This finding could be useful in designing a new type of feedback to rethink assessment technologies and public speaking.

ACKNOWLEDGMENTS

REFERENCES

1. Roghayeh Barmaki. 2016. Improving Social Communication Skills Using Kinesics Feedback. In *2016 CHI Conference Extended Abstracts*. ACM, 86–91.
2. Geoffrey W Beattie and Brian L Butterworth. 1979. Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech* 22, 3 (1979), 201–211.
3. Emmanuel Biau and Salvador Soto-Faraco. 2013. Beat gestures modulate auditory integration in speech perception. *Brain and language* 124, 2 (2013), 143–152.
4. Christopher M Bishop. 2006. *Pattern recognition and Machine Learning*. Vol. 128. Springer.
5. Paul Boersma and others. 2002. Praat, a system for doing phonetics by computer. *Glot international* 5, 9/10 (2002), 341–345.
6. Mark Bubel, Ruiwen Jiang, Christine H Lee, Wen Shi, and Audrey Tse. 2016. AwareMe: Addressing Fear of Public Speech through Awareness. In *CHI Conference Extended Abstracts*. ACM, 68–73.
7. Rich Caruana and Dayne Freitag. 1994. Greedy Attribute Selection.. In *International Conference on Machine Learning (ICML)*. 28–36.
8. Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2 (2011), 27:1–27:27. Issue 3. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
9. Lei Chen, Gary Feng, Jilliam Joe, Chee Wee Leong, Christopher Kitchen, and Chong Min Lee. 2014. Towards automated assessment of public speaking skills using multimodal cues. In *International Conference on Multimodal Interaction (ICMI)*. ACM, 200–203.
10. François Chollet. 2015. Keras. https://github.com/fchollet/keras. (2015).
11. Martin Corley and Oliver W Stewart. 2008. Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass* 2, 4 (2008), 589–602.
12. Ionut Damian, Chiew Seng Sean Tan, Tobias Baur, Johannes Schöning, Kris Luyten, and Elisabeth André. 2015. Augmenting social interactions: Realtime behavioural feedback using social signal processing techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 565–574.
13. Anthony Steven Dick, Susan Goldin-Meadow, Uri Hasson, Jeremy I Skipper, and Steven L Small. 2009. Co-speech gestures influence neural activity in brain regions associated with processing semantic information. *Human brain mapping* 30, 11 (2009), 3509–3526.
14. Ronald A Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics* 7, 2 (1936), 179–188.
15. Stanley Grand. 1977. On hand movements during speech: Studies of the role of self-stimulation in communication under conditions of psychopathology, sensory deficit, and bilingualism. In *Communicative structures and psychic structures*. Springer, 199–221.

16. James A Hanley and Barbara J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 1 (1982), 29–36.

17. Jinni A Harrigan. 1985. Self-touching as an indicator of underlying affect and language processes. *Social Science & Medicine* 20, 11 (1985), 1161–1168.

18. Bill Hoogterp. *Your Perfect Presentation: Speak in Front of Any Audience Anytime Anywhere and Never Be Nervous Again*.

19. Klaus Krippendorff. 2004. *Content analysis: An introduction to its methodology*. Sage.

20. Kris Luyten, Donald Degraen, Gustavo Rovelo Ruiz, Sven Coppers, and Davy Vanacken. 2016. Hidden in Plain Sight: an Exploration of a Visual Language for Near-Eye Out-of-Focus Displays in the Peripheral View. In *CHI Conference on Human Factors in Computing Systems*. ACM, 487–497.

21. David McNeill. 2008. *Gesture and thought*. University of Chicago Press.

22. Sandra Merlo and Letıcia Lessa Mansur. 2004. Descriptive discourse: topic familiarity and disfluencies. *Journal of Communication Disorders* 37, 6 (2004), 489–503.

23. Joseph Murphy. 2012. *The power of your subconscious mind*. Courier Corporation.

24. Iftekhar Naim, M Iftekhar Tanveer, Daniel Gildea, and others. 2015b. Automated Analysis and Prediction of Job Interview Performance. *arXiv preprint arXiv:1504.03425* (2015).

25. Iftekhar Naim, M Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2015a. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *Automatic Face and Gesture Recognition (FG)*, Vol. 1. IEEE, 1–6.

26. Anh-Tuan Nguyen, Wei Chen, and Matthias Rauterberg. 2012. Online feedback system for public speakers. In *E-Learning, E-Management and E-Services (IS3e), 2012 IEEE Symposium on*. IEEE, 1–5.

27. Sharon Oviatt. 1995. Predicting spoken disfluencies during human–computer interaction. *Computer Speech & Language* 9, 1 (1995), 19–35.

28. F. Pedregosa and others. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

29. AKM Mahbubur Rahman, Md Iftekhar Tanveer, Asm Iftekhar Anam, and Mohammed Yeasin. 2012. IMAPS: A smart phone based real-time framework for prediction of affect in natural dyadic conversation. In *Visual Communications and Image Processing (VCIP), 2012 IEEE*. IEEE, 1–6.

30. Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. 2009. Face alignment through subspace constrained mean-shifts. In *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 1034–1041.

31. Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. 2011. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision* 91, 2 (2011), 200–215.

32. Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. 2013. Real-time human pose recognition in parts from single depth images. *Commun. ACM* 56, 1 (2013), 116–124.

33. Alex J Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and computing* 14, 3 (2004), 199–222.

34. Vikrant Soman and Anmol Madan. 2009. Social signaling: Predicting the outcome of job interviews from vocal tone and prosody. In *IEEE International Conference on Acoustic, Speech and Signal Processing*. Citeseer.

35. M Iftekhar Tanveer, Emy Lin, and Mohammed Ehsan Hoque. 2015a. Rhema: A Real-Time In-Situ Intelligent Interface to Help People with Public Speaking. In *20th International Conference on Intelligent User Interfaces*. ACM, 286–295.

36. M Iftekhar Tanveer, Ji Liu, and M Ehsan Hoque. 2015b. Unsupervised Extraction of Human-Interpretable Nonverbal Behavioral Cues in a Public Speaking Scenario. In *23rd Annual ACM Conference on Multimedia*. ACM, 863–866.

37. M. Iftekhar Tanveer, Ru Zhao, Kezhen Chen, Zoe Tiet, and Mohammed Ehsan Hoque. 2016. AutoManner: An Automated Interface for Making Public Speakers Aware of Their Mannerisms. In *21st International Conference on Intelligent User Interfaces (IUI '16)*. ACM, New York, NY, USA, 385–396.

38. Theano Development Team. 2016. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints* abs/1605.02688 (May 2016). http://arxiv.org/abs/1605.02688

39. Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.

40. Toastmasters International. 2011. Gestures: Your body speaks. Online Document. Available at http://web.mst.edu/~toast/docs/Gestures.pdf. (2011).

41. Christopher Turk. 2002. *Effective speaking: Communicating in speech*. Routledge.

42. Vladimir Vapnik and Alexey Chervonenkis. 1964. A note on one class of perceptrons. *Automation and remote control* 25, 1 (1964).

43. Timothy D Wilson. 2004. *Strangers to ourselves*. Harvard University Press.

44. Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America* 123, 5 (2008), 3878.