AutoManner: An Automated Interface for Making Public Speakers Aware of Their Mannerisms

M. Iftekhar Tanveer[1], **Ru Zhao**[2], **Kezhen Chen**[3], **Zoe Tiet**[4], **Mohammed (Ehsan) Hoque**[5]
Rochester Human-Computer Interaction (ROC-HCI), University of Rochester
{[1]itanveer, [5]mehoque}@cs.rochester.edu
{[2]rzhao2, [3]kchen33, [4]ztiet}@u.rochester.edu

ABSTRACT

Many individuals exhibit unconscious body movements called mannerisms while speaking. These repeated changes often distract the audience when not relevant to the verbal context. We present an intelligent interface that can automatically extract human gestures using Microsoft Kinect to make speakers aware of their mannerisms. We use a sparsity-based algorithm, *Shift Invariant Sparse Coding*, to automatically extract the patterns of body movements. These patterns are displayed in an interface with subtle question and answer-based feedback scheme that draws attention to the speaker's body language. Our formal evaluation with 27 participants shows that the users became aware of their body language after using the system. In addition, when independent observers annotated the accuracy of the algorithm for every extracted pattern, we find that the patterns extracted by our algorithm is significantly ($p < 0.001$) more accurate than just random selection. This represents a strong evidence that the algorithm is able to extract human-interpretable body movement patterns. An interactive demo of AutoManner is available at **`http://tinyurl.com/AutoManner`**.

Author Keywords

Public Speaking; Body Language; Interface Design

ACM Classification Keywords

I.5.0 Pattern Recognition: General; H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Have you ever felt unaware of your body language while giving a public speech? What did you do with your hands? Did you move around while speaking or stand still? How did you use your gestures to emphasize a point that you wanted to make?

Our conscious mind can process only 40 bits of information per second [22, 38]. As a result, we become easily overwhelmed in public speaking while thinking about what to say next. In this situation, many of our innate functions are delegated to the subconscious mind, which can process up to 4 million bits per second. This often results in unconscious and possibly repetitive patterns of movements. These movements include gripping or leaning, tapping fingers, whole body movements (such as rocking, swaying, pacing), jingling pocket change, adjusting hair or clothing, etc. The Toastmasters group [36] refers to these as *mannerisms*. These can significantly distract the audience.

In this paper, we present—AutoManner—an interface to automatically extract and visualize the repetitive patterns of the speaker's movements during public speaking. AutoManner captures and analyzes body language using a Microsoft Kinect [39] depth sensor to make speakers aware of their mannerisms. We implement a subtle feedback technique that involves pinpointing frequent movements and asking questions about their meanings. In our experiment with 27 participants, the users reported that the use of AutoManner made them aware of their mannerisms.

Most of the existing systems [27, 24, 6, 33] that assess the quality of public speech utilize a supervised classification approach. In supervised methods, the system either assigns the speakers to a predefined category or a number indicative of the quality of the speech. Although this *supervised approach* is beneficial for ontological categorization, it may not provide any insights about the strengths and weaknesses of an individual. For example, it usually does not offer any insight as to why particular body language is rated poorly, or which patterns in their body movement were not effective. In addition, a supervised approach is unable to detect idiosyncratic or unexpected body movements, as it is difficult to define all possible patterns that a person may show in a given context.

One possible way to solve this problem is to share the videos with public speaking experts and get subjective feedback. Challenges include identifying those experts and paying for their time. Another possibility is to obtain cheap micro-level annotations in the cloud using crowdsourcing [26]. However, crowdsourcing poses an inherent threat to the speaker's privacy. Speakers may not feel comfortable about real people viewing their videos and judging their speaking performance and body language. In this paper, we address this challenge by developing a fully automated framework that allows users to obtain feedback while being in complete control of their data.

In one of our previous projects [35], we proposed a sparsity-based algorithm to extract repetitive body movements (behavioral cues). In the AutoManner interface, we use this

IUI 2016, March 7–10, 2016, Sonoma, CA, USA.

DOI: http://dx.doi.org/10.1145/2856767.2856785

algorithm to automatically extract and show these repetitive movements to the speaker.

However, not every repetitive movement is bad. Specific gestures can be distracting (mannerism) or meaningful depending on how cohesive it is with the verbal content [13]. Currently, the proposed interface cannot determine the context of the verbal content. Nevertheless, it makes the speakers aware of their body language by highlighting the repeated body movements and asking questions about them. Reflection through answering questions about body language is a subtle way to show the speaker's idiosyncrasies and mannerisms. In our experiments, we found evidence supporting this design, as the participants reported becoming self-aware of their body language. A similar effect of awareness might also be achieved if one carefully watches and annotates the entire video. AutoManner automatically highlights the regions of repetitive body movements, thus eliminating the tedious manual observation and annotation process.

We claim the following contributions in this paper:

- We develop an intelligent interface that can automatically extract repetitive patterns of body language and visualize them to make public speakers aware of their body language.
- We formally evaluate the interface with 27 public speakers. They self-reported to become aware of their mannerisms after using our interface.
- Our experiment provides strong evidence ($p < 0.001$) that the algorithm we used is able to extract more accurate and human-interpretable body movement patterns than just random sampling.
- To the best of our knowledge, this is the first attempt to automatically analyze mannerisms in public speaking.

LITERATURE REVIEW

Nonverbal behavior (e.g. body language, vocal prosody, facial expressions, etc.) is an important modality for interpersonal communication. It is shown to be predictive of human communication skills in many different domains—for instance, patient satisfaction [11], analysis of persuasiveness [30], dating [28], job interviews [2, 23], etc. Body language is an important component of nonverbal behavior [16]. Cognitive neuroscientist Beatrice de Gelder [10], after an extensive literature review, concluded that body language is as reliable a metric as facial expressions.

Constituents of a Good Public Speech

Both verbal and nonverbal behaviors are important factors in good public speaking [32]. Strong speakers communicate ideas through a delicate interplay between verbal content and body language (body postures, head and hand movements, etc.). Schreiber et al. [29] analyzed the characteristics of good public speeches in order to develop a rubric for judging public speaking competency. Their work substantiates the importance of congruence between verbal and nonverbal factors during public speaking.

Experts on public speaking [13, 9, 18, 36] usually stress the importance of coherence in verbal content and body language. Hoogterp mentioned in his book [13] that body language should be synchronized with a speech for effective nonverbal communication. In addition, hand gestures should convey the same meaning as the verbal content. The Toastmasters group [36] pointed out that eliminating distracting mannerisms was necessary for conveying spontaneous and genuine feelings through body language.

Automated Assessment of Public Speaking Competence

A considerable amount of work has been done to model and automatically predict the performances of public speaking. Pfister and Robinson [27] proposed a real-time system to classify affective states and to assess public speaking skills. They extracted prosodic features such as pitch, intensity, and *MFCC* to classify a speaker's affective states into one of nine discrete categories (absorbed, excited, interested, joyful, opposed, stressed, sure, thinking, unsure) using a Support Vector Machine (SVM). Similarly, they classified public speaking skills into one of six discrete classes (clear, competent, credible, dynamic, persuasive, and pleasant).

Nguyen et al. [24] proposed an online feedback system that provides feedback about a speaker's body language on a scale of five degrees, from bad to excellent. They used a Kinect skeleton tracker to track the postures and gestures of the speaker. They then used a nearest-neighbor classifier to compare those movements to a set of predefined templates of postures and gestures in order to determine the quality of the speaker's speech. This method cannot account for completely novel body language. In addition, this method is not suitable for assessing cases where a gesture is meaningful in one context but not meaningful in another. Chen et al. [6] described 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) to predict a continuous score indicating public speaking performance. They claimed a correlation coefficient of 0.38 to 0.48 with the manually annotated ground truth.

All the systems discussed so far utilize supervised classification/regression approaches. This may be an artifact of similar trends in human action/activity recognition literature [1, 7, 19, 37]. Recently, some work has been done on unsupervised analysis of human action detection. Niebles et al. [25] proposed a method of unsupervised learning of human action categories using a Latent Dirichlet Allocation model. Zhou et al. [40] proposed Aligned Cluster Analysis (ACA), which detects patterns in signals using k-means clustering and dynamic time warping. Tanveer et al. [35] proposed a sparse coding based approach for detecting common behavioral cues for body language. In this paper, we use the method mentioned in Tanveer et al. for capturing the fidgeting patterns and the mannerisms of public speakers.

Design of Effective Feedback

Feedback is an important component for any nonverbal skills assessment and training system. The predicted skill levels

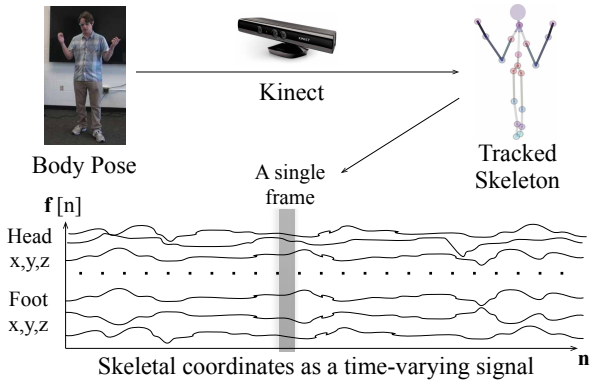

Figure 1: Steps for capturing the Motion Capture (MoCap) signal.

and useful recommendations should be effectively and intuitively conveyed to the users. A significant amount of work has already been done on designing useful feedback systems. For instance, Hoque et al. [14] showed that it is possible to improve job interview performances by reviewing one's own videos augmented with raw features such as smile, head movements, speech, and prosody. They used a virtual avatar to practice the job interviews. Tanaka et al. [33] also employed virtual characters for practicing social skills for people with Autism Spectrum Disorder (ASD). Tanveer et al. [34] designed a system to provide live feedback on prosodic behavior using Google Glass. They utilized *secondary display* phenomenon for minimizing distractions during the speech.

Chollet et al. [8] performed extensive analysis on learning strategies and usefulness of a virtual audience in public speaking training. They analyzed the efficacy of an interactive learning framework (Cicero [3]) where speakers can practice speaking in a safe and engaging environment with a virtual audience. The virtual audience can provide nonverbal feedback to signal elevated attention, rapport, lack of interest, or disagreement. They show that the interactive virtual audience results in significant progress in perceived attention and combined improvement compared to the traditional ways of providing feedback.

TECHNICAL DETAILS

Many public speakers display their own idiosyncratic body language, which makes it difficult to compile an exhaustive list of all movement patterns. Without an exhaustive list, we cannot train a supervised classification system to detect patterns of body movements. To address this challenge, we have developed an algorithm to extract common body movements automatically—without any human supervision on segmenting or labeling of a training dataset. In this section, we describe this method.

Data Capture

In order to sense the movements of the speakers, we record the speaker's activity using a Kinect [39] depth sensor. Kinect uses an infrared projector and sensor arrangement to record depth images of the region in front of the device. We use a skeleton tracker [31] to extract the 3D coordinates of a person's body from the depth images. The skeleton tracker can track twenty joint locations of the person's body, as shown in Figure 1. We refer to the time-sequence of all the joint locations as a Motion Capture (MoCap) signal.

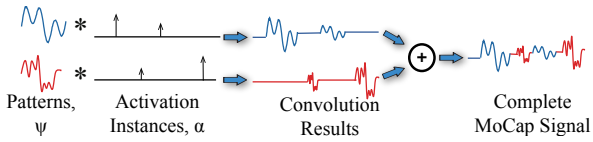

Figure 2: A simplified illustration of MoCap signal model. Only one component is shown ($k = 1$).

Preprocessing

The motion capture signal is a high dimensional vector-valued signal. At each time-instance, the signal consists of 60 (20 joints $\times$ 3 components—x, y, z) components. Due to the logistics of the experimental setup (i.e. to limit the total study time) we needed to reduce the dimensionality of the input signal. We did it by applying Principal Component Analysis (PCA) [15] at the frame level. The number of basis vectors for PCA was determined through a heuristic evaluation over a sample dataset. We kept the basis vectors corresponding to the k largest eigenvalues that account for 99% of the total variance of the signal. The value of k usually varies from five to 16, depending on the contents of the signal. PCA helps to reduce the noise in the MoCap signal by eliminating low energy principal components.

Extracting the Behavioral Cues

Behavioral cues are defined as small meaningful repetitive patterns in a person's gestures, posture, touching behavior, facial expressions, eye behavior, vocal behavior, etc. [16, 37] In this paper, we disregard the aspect of "meaning" for behavioral cues and use it interchangeably as "body movement patterns." In our prior research [35], we proposed an unsupervised algorithm to extract human-interpretable patterns of body movements from MoCap signals. Our algorithm was inspired by the *Shift Invariant Sparse Coding (SISC)* algorithm, proposed by Morup et al. [21]. In this project, we have developed an interface to display these patterns to the speakers as potential mannerism candidates. We briefly describe the technique in this section. Please refer to Tanveer et al. [35] for an elaborate discussion of this algorithm.

Mathematical Model

Let us assume that the MoCap signal, $\mathbf{f}[n]$, has k components, and the length of the signal is N. The behavioral cues are manifested in the signal as a specific pattern of variations. The patterns can appear at any location in the signal. Let us assume that there are D unique patterns and we denote any d^{th} pattern as $\psi_d[m]$. We also denote the signal representing the activation instances of the d^{th} pattern as $\alpha_d[n]$, which is just a collection of impulse functions, as shown in Figure 2. We model the MoCap signal through a superposition of all the patterns repeated at the corresponding activation instances. It

can be described mathematically as in Equation (1).

$$\begin{aligned}\mathbf{f}_m[\mathrm{n}] &= \sum_{d=0}^{D-1} \alpha_d[n] * \psi_d[m] \\ &= \sum_{d=0}^{D-1} \sum_{u=0}^{N-1} \alpha_d[u]\psi_d[n-u].\end{aligned} \tag{1}$$

Here, the asterisk ($*$) symbol denotes the convolution operation. Convolution of any signal with an impulse function results in a time-shift of the signal into the time-location of the impulse function.

Problem Formulation

We extract the behavioral cues by estimating ψ and α that minimize the total squared error between our MoCap model and the actual MoCap signal. However, without any suitable constraint, this minimization problem has many solutions. In order to get a unique solution, we enforce a sparsity penalty over the activation instances, α. In other words, we assume that any particular pattern occurs only sporadically over the signal. This assumption is reasonable because, in reality, it is not possible for a particular pattern to occur densely without distorting itself. The mathematical form for the optimization problem is shown in Equation (2).

$$\begin{aligned}\hat{\psi}[m], \hat{\alpha}[n] = \underset{\psi,\alpha}{\operatorname{argmin}} \underbrace{\frac{1}{2} \left\|\mathbf{f}[n] - \mathbf{f}_m[n]\right\|^2}_{P(\psi,\alpha)} + \lambda \left\|\alpha\right\|_1 \\ \text{s.t.} \quad \left\|\psi\right\|_F^2 \leq 1 \quad \text{and,} \quad \forall_n \alpha[n] \geq 0.\end{aligned} \tag{2}$$

Here, the $P(\psi, \alpha)$ term represents the squared error. The ℓ-1 norm of α is the term for enforcing sparsity and λ is a multiplier controlling the relative proportion for the error term and the ℓ-1 norm. The constraint $\|\psi\|_F^2 \leq 1$ makes sure that ψ cannot grow arbitrarily large. Finally, the constraint $\alpha[n] \geq 0$ ensures that the activation instances are non-negative. This is important to make sure that the extracted patterns are not upside down.

Algorithm

The objective function in this optimization problem is non-convex, in general. However, it is convex over one of the parameters (α or ψ) when another parameter is fixed. Therefore, we can solve it using an *Alternating Proximal Gradient Descent* approach. The complete procedure is shown in Algorithm 1. We alternatively keep one parameter fixed and update another, and vice versa. As the error reduces at each iteration of the gradient descent process, the algorithm is guaranteed to converge.

However, we cannot calculate the gradient of the objective function at all the points. The ℓ-1 norm of α is non-smooth, and thus non-differentiable. We use an *Iterative Shrinkage Threshold Algorithm (ISTA)* [5] to solve this problem. We use the gradients of the smooth part (i.e. $P(\psi, \alpha)$) to update the parameters. Then, at each iteration, we apply a *shrink* operation over α to enforce sparsity. We also project α to the set of non-negative numbers to enforce the non-negativity constraint. The gradients of $P(\psi, \alpha)$ with respect to ψ and α

Algorithm 1: Learning the Behavioral Cues

Input: $\mathbf{f}[n]$, M, D and λ
Output: ψ, α
Initialize;
$i \leftarrow 0$;
$\alpha \leftarrow 0$, $\psi \leftarrow$ random;
while *not Converge* **do**
  Update ψ;
  reconstruct $\mathbf{f}_{model} \leftarrow \sum_{d=1}^{D-1} \alpha_d * \psi_d$;
  calculate $\nabla_\psi P$ using $\mathbf{f}, \mathbf{f}_{model}$ and α [Eq. (3)];
  $\psi^{(i+1)} \leftarrow \mathbf{project}(\psi^{(i)} - \gamma_\psi \nabla_\psi P)$;
  Update α;
  reconstruct $\mathbf{f}_{model} \leftarrow \sum_{d=1}^{D-1} \alpha_d * \psi_d$;
  calculate $\nabla_\alpha P$ using $\mathbf{f}, \mathbf{f}_{model}$ and ψ [Eq. (4)];
  $\alpha^{(i+1)} \leftarrow \mathbf{shrinkandproject}(\alpha^{(i)} - \gamma_\alpha \nabla_\alpha P)$;
  $i \leftarrow i + 1$

are shown in equations (3) and (4) respectively.

$$\frac{\partial P}{\partial \psi_{d',k'}[m']} = \sum_{n=0}^{N-1} \{f_{\text{model},k'}[n] - f'_k[n]\}\alpha_{d'}[n - m'] \tag{3}$$

$$\frac{\partial P}{\partial \alpha_{d'}[n']} = \sum_{k=0}^{K-1} \sum_{n=0}^{N-1} \{f_{\text{model},k}[n] - f_k[n]\}\psi_{d',k}[n - m'] \tag{4}$$

The **shrinkandproject** method is shown in equation (5).

$$\begin{aligned}&\alpha[n] \leftarrow \text{sgn}(\alpha[n])\text{max}(0, |\alpha[n]| - \gamma\lambda) \quad \forall_{0 \leq n < N} \\ &\alpha[n] \leftarrow max(0, \alpha[n]) \quad \forall_{0 \leq n < N}\end{aligned} \tag{5}$$

Please note that, although the algorithm simultaneously solves two convex problems, the whole objective function is not convex, in general. Therefore, it is not guaranteed that the algorithm will reach the global optima when it converges—it may settle upon a local optima. We randomly initialize the algorithm, run it multiple times, and choose the lowest value to make it more likely to reach the global optima. The length of each pattern is heuristically set to two seconds in this application.

Speedup Techniques

While designing the user studies for the system, we needed to extract the behavioral cues between two consecutive speeches by a speaker. This situation imposed additional constraints on the convergence time for the algorithm. We wanted to ensure that it does not take more than five minutes to converge for a three-minute MoCap sequence. To achieve this, we reduced the dimensionality of the signal using PCA, as described before. We also set the number of patterns to be extracted at five, as makes the program linearly slower. However, even with these settings, the algorithm took about twenty minutes to converge, which is unacceptable, according to the study design.

We choose the step-sizes (γ_ψ and γ_α) using a *bold driver* [4] method to further speed up the algorithm. Choice of step size

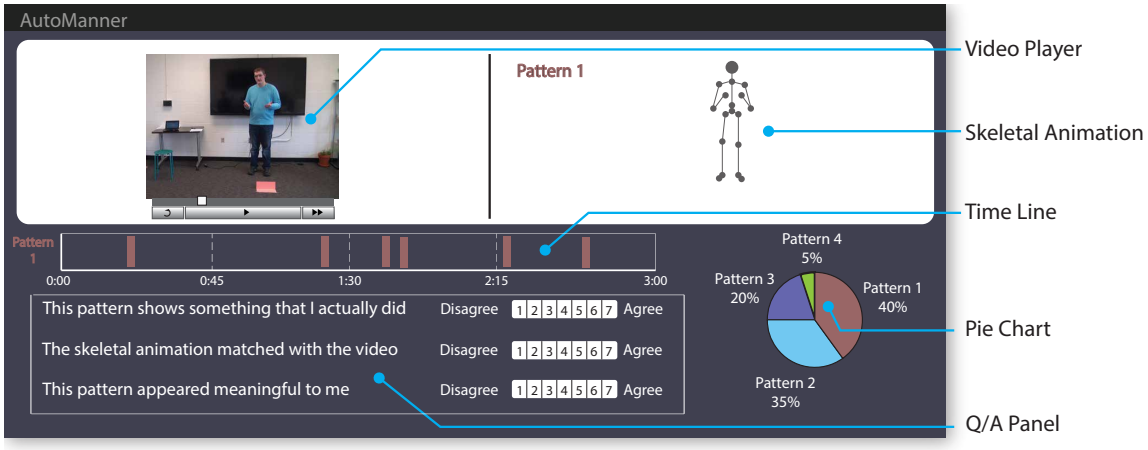

Figure 3: A screenshot of the user interface designed to analyze body language. A live sample is available at: **http://tinyurl.com/AutoManner**

is important, as too small a value will make the algorithm slow to converge. On the other hand, large step sizes will make it diverge from the optimum. In the bold driver method, the step-sizes are slowly increased (by five percent) at each iteration that reduces the value of the objective function. If the value of the objective function increases at any iteration, the step size is largely penalized (by 50 percent decrease) because it indicates the step size is too large. We find in empirical tests that bold driver is faster to perform than a typical *Line Search* approach, described by Tanveer et al. [35]. In addition, it takes a lower number of iterations than a constant step size. We reduced the convergence time to below five minutes by using the bold driver approach. Finally, we use simultaneous parallel jobs to rerun the algorithm and choose the results from the best performing job. We use a cluster computing environment for the parallel execution of the jobs.

DESIGN OF THE INTERFACE

The SISC algorithm described above extracts the common behavioral cues related to the speaker's body language. However, the appropriateness of the body language is largely dependent on the context of the speech—which is difficult to automatically assess. Therefore we design a user interface to make the speakers think about their own body language using a question-answer technique. We automatically extract the body movement patterns and ask questions to induce thoughts about the meaning of the cue in context of verbal content. We hypothesize that the process of thinking about a specific pattern of body movement will make them aware of their body language.

There are two main components in the interface. The first component shows the video of the speech and asks some questions about the overall quality of the speech. The video could be played normally, or fast-forwarded to quickly review the body movements in the entire speech. The second component of the interface is designed to make the speaker review and think about the extracted behavioral cues (i.e. patterns of body movements) one at a time. The speaker reviews the extracted behavioral cues and answers three questions related to each of them. This is shown in Figure 3. The interface also incorporates a video player for watching the whole speech. In addition, it contains a panel for a skeletal animation, a timeline, a pie chart, and a Q/A section. The *skeletal animation* in the upper right side of the interface represents a specific pattern of body movement (i.e. behavioral cue). This is the ψ parameter as denoted in Algorithm 1. The *timeline* shows the corresponding activation instances (α) for that specific pattern. The users can click on the time instances to play the region of the video where the pattern is activated. This allows the user to judge the verbal context for that specific behavioral cue. The *pie chart* gives a general overview to the relative proportion of the extracted patterns. The algorithm extracts the top five patterns of body movements. However, there might be less than five patterns depending on how diverse the speaker's body language is. In case the number of extracted pattern become less than three, the interface will display the following warning: "You did not move enough. For good body language you should move more and move purposefully".

A question-answer panel asks the users to rate three statements related to each pattern shown in the interface. These are listed in the "Pattern Specific Ratings" row of Table 1. The first statement makes the speaker observe the skeleton movements. The second statement makes them analyze the occurrences of the patterns in the timeline. The third statement makes them judge how coherent of the patterns are with the context of the speech. As we discussed before, the primary objective of these questions are to make the users think about the body language. Observing the skeletal movements will make them aware of a specific movement in their body. Besides, analyzing the time-instances of the pattern occurrence will provide a good idea on the context of the speech. We prepared a video tutorial[1] that explains the use of the interface. The participants of the user study watched it before their first interaction with the interface.

[1] **http://tinyurl.com/AutoMannertutorial**

Figure 4: Lab settings for the public speaking study

RESEARCH QUESTIONS

We wanted to answer the following research questions related to AutoManner in this paper.

- How useful is the interface? Do the users report becoming more self-aware of their body language?
- How well does the developed algorithm accurately identify the relevant behavioral patterns? If independent observers (e.g. workers of Mechanical Turk) annotate the accuracies of each time instance, is there any significant difference of accuracy between the treatment and the placebo patterns?
- Does the AutoManner interface help speakers improve their body language?

These questions are of different levels of difficulty which we attempted to answer from various perspectives using both quantitative and qualitative metrics. We decided to answer the first question from the user's point of view. We asked the users to evaluate the usefulness by annotating their opinions about a few statements. We also asked their subjective opinions in a free form question-answer session to gain more insight about AutoManner's efficacy. In the second question, we wanted to know the algorithm's accuracy. This could have been objectively answered by calculating error between the original signal and the extracted patterns. However, we are more interested on human-perception of the algorithmic efficiency, rather than just an error rate. Therefore we recruited Mechanical Turk™ workers to observe and annotate the correctness of time instances that the algorithm detected. The turkers annotated the meaningfulness of the patterns as well. In the third question, we attempted to seek any improvement in the participants body language.

STUDY DESIGN

The design of the formal user study is focused on answering the research questions described above. Figure 5 illustrates an outline of the study. We asked the participants to deliver three speeches, each of which is three minutes in duration. After each speech, the participants rated their own speeches. Once they finished self-rating, they started interacting with the interface. The interaction involved observing each pattern of body movements and answering a few questions. Once finished the interaction, the participants rated the whole interface for its the usefulness. The same process continues for the second speech. No interface was shown after the third speech—the participants only rated themselves. The measures are shown in Table 1. All these statements were rated in a 7-point Likert scale [17].

Self-Ratings (for each speech)	Overall, I'm happy with the quality of my speech. I was purposely moving about. I was using a variety of gestures. My gestures were appropriate with the speech.
Pattern Specific Ratings (for each pattern)	This pattern shows something I actually did. The skeleton animation matched with the video. This pattern appeared meaningful to me.
Interface Ratings (Asked after reviewing all the patterns)	The feedback was very helpful. The feedback made me aware of my body language. This feedback made me aware of at-least one gesture that I make too frequently. I will use the system if it is available

Table 1: List of measures rated by the participants. All of these statements were rated in a 7-point Likert scale where higher values indicate better.

Baseline, Treatment, and Placebo

Among the three speeches, the first one was delivered without any interaction with our interface. Performance in this speech is considered as a baseline in the study. In order to evaluate the correctness of the algorithm, we prepared two different versions of the feedback interface. In one version, the interface showed the real patterns and time-instances extracted by the algorithm. We refer to this as the *treatment* interface. In another version, we faked the patterns by randomly sampling 2-second windows from the MoCap signal. The time-instances of these patterns were also randomly selected in the time-line. This version of the interface is referred as the *placebo*. In order to collect sufficient data, we made each participants to interact with both the placebo and the treatment—however, their order was counter-balanced. Participants with odd ID-number saw the placebo interface first, whereas even ID holders saw the treatment interface first. The participants did not know their ID number and the fact that there were two different versions of the interface.

Demographics, and Lab Settings

Twenty-seven people participated in the user study. Among them, 14 were female and 13 male. All of participants were undergraduate students at the University of Rochester. They declared themselves as native speakers of English. We recruited the participants by putting flyers around the campus and also by posting invitations in the student's Facebook pages. The participants received ten dollars in the form of Amazon gift cards for participating in the study. The study took approximately forty minutes to finish.

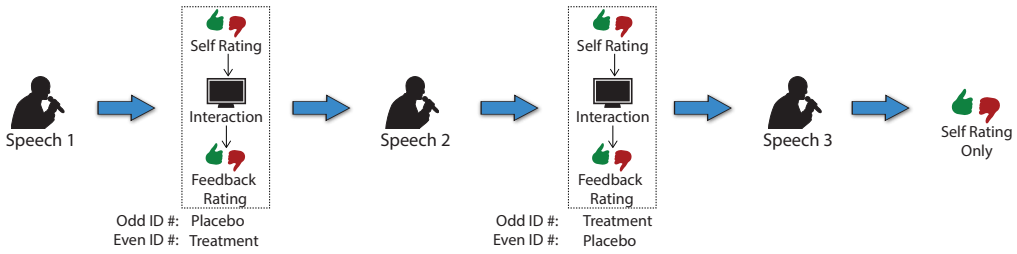

Figure 5: Design of the user-study to answer the research questions

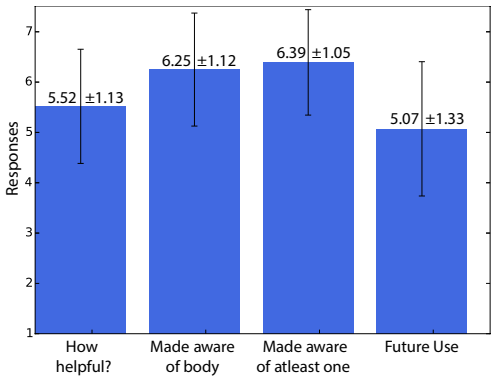

Figure 6: Interface related responses by the participants. The error-bars show one standard deviation from the mean. The dotted line shows the maximum value of Likert Scale.

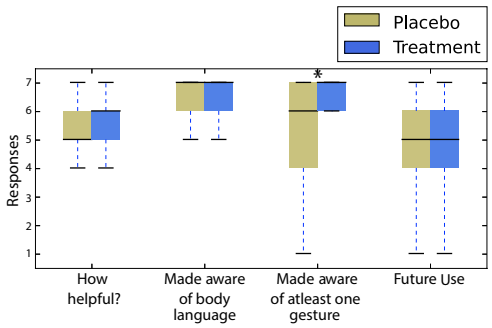

Figure 7: Box-Whisker plot for interface related responses. The ratings are grouped in placebo and treatment interfaces.

During the study, we video-recorded the speeches using a Canon Rebel T3 DSLR camera. We captured the full body motion capture (MoCap) as well, using a Kinect depth sensor and Microsoft SDK. We used a custom arrangement for time-synchronized records of the video and the MoCap signal. A picture of the recording arrangement is shown in Figure 4. The speakers chose three topics from a list of sample topics (e.g. favorite pastime, favorite book/movie/superhero, etc.) that we supplied for convenience. However, they were free to choose any other topic of their interest. The topics were decided at least one day ahead to allow the participants to prepare for the speech.

Finally, we randomly selected 7 participants to conduct short one-to-one interviews at the end of study. In this interview, we asked for their subjective opinions about the interface and audio-recorded their responses. This interview was done only after we finished the study. We asked if the participants liked the interface or not and the reasons for liking or disliking it. We also asked if they could realize that there were two different versions of the interface—one real (treatment) and another fake (placebo). In addition, we asked them to guess which one was real and which one was fake. Finally, we asked if the interface helped them become aware of their body languages or not.

RESULTS

In order to evaluate the usefulness of the interface, we analyze the participants' responses to the interface related measures (i.e. the third row in Table 1) in our study. These responses are recorded in a 7-point Likert scale. We compute the average of the aggregated responses from the real feedback only. In other words, the average is calculated by considering the responses after the first speech for the participants with even ID number, and responses after the second speech for participants with odd ID number. The average responses are shown in Figure 6. The error bars represent one standard deviations from the average. It is evident from the figure that most of the responses are more than 5 in the 7-point Likert scale. The participants provided higher ratings in the measures related to the awareness of body language. This represents the fact that AutoManner can successfully make the users aware of their body language.

We perform a group-wise analysis in order to observe the differences in the participants' responses in placebo and treatment interface. We group all the responses after interaction with the placebo interface as "*placebo*"; and that after the treatment interface as "*treatment*". A Box-Whisker plot for these two groups is shown in Figure 7. In this plot, the box represents the lower and upper quartile of the data. The horizontal bar inside the box represents the median. The whiskers represent the range of the data. We also conducted a Wilcoxon-Mann-Whitney Rank Sum test [12] to measure the statistical significance of differences between the groups. Throughout the paper, we use single, two, and three asterisk symbols to represent a rejection of the null hypothesis with 0.05, 0.01, and 0.001 significance levels respectively. Notice that there is a statistically significant difference ($p < 0.05$) in the responses for the third statement. It is clear from the figure that the responses for the placebo group varies widely in comparison to the treatment group. Responses for other

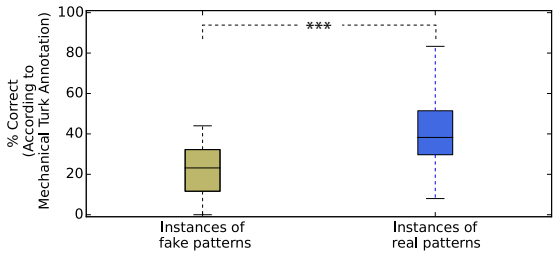

Figure 8: Box-Whisker plot for the percentages of instances for real and fake patterns. The three asterisk symbol represents a statistical significance with $p < 0.001$

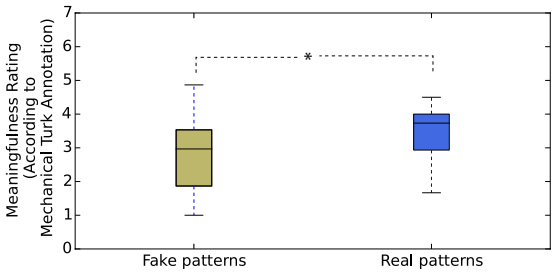

Figure 9: Ratings for "meaningfulness" of the patterns as annotated by the turkers. $p < 0.05$

measures are not statistically significant. This result makes sense considering the fact that the third statement is directly related to the parameter that we manipulated in the placebo and treatment interface.

In the second research question, we want to evaluate the efficacy of the algorithm from human perspective. On this regard, we require annotations on the accuracy of each time-instances. However, this is a huge task as there are as many as 50 time-instances for each pattern and 5 patterns per video. It was impractical to ask the participants to annotate the accuracy for each instance in the video. We solve this problem by recruiting thirty workers in the Mechanical Turk[2] website. In order to ensure high quality in the answers, we accept the turkers who completed at-least 1000 tasks with 99% acceptance rate. In addition, we perform a qualifying round; where we manually selected 30 turkers based on their performance on annotating a ground truth. These filtering and qualification round techniques were performed in light to the work of Mitra et al. [20].

We use an interface similar to AutoManner for the turkers to view the videos, patterns, and time-instances. The only difference between the mechanical turk interface and the original AutoManner is in the questions asked through the interface. This interface requires the turkers to annotate the answer to the following question for each time-instance: "*Does the skeletal animation match with the video at this time-instance?*". The turkers respond either yes or no. In addition, while evaluating the patterns, they annotate if they agree to the following statement: "*This body movement pattern conveys a meaning*". The response is recorded in a 7-point Likert scale. Three different turkers annotate the same data-point. We consider the mode (i.e. majority vote) of the responses as the turker response for a particular data point.

For the placebo videos, we extract the corresponding real patterns using the SISC algorithm. These real patterns were also shown to the turkers and received annotations. In our experiment, the turkers did not have any knowledge about the treatment or placebo groups of the patterns. For each pattern, we calculate the percentages of the time-instances that the turkers marked as correct. We represent the data in two groups—one for the fake patterns (placebo), another for the corresponding real patterns from the same videos. Figure 8 represents a box-and-whisker plot of these accuracies. It is evident in the figure that there is a clear difference in the perceived accuracy between the instances of the fake patterns and the real patterns. The mechanical turk workers were not aware of the differences in the patterns at all. In addition, different workers annotated different instances of the patterns. Despite this, we observe a statistically significant difference between the time-instances of real patterns and fake patterns. This is a strong evidence that the SISC algorithm performs significantly well in detecting and localizing the patterns. As the correctness of the time instances were marked by human annotators, this result also represents that the algorithm is able to extract human-interpretable body movement patterns.

We illustrate the turker's responses over the statement—"This body movement pattern conveys a meaning" in Figure 9. It is evident from this figure that the difference of responses between the fake and the real patterns is not as significant as the differences in their time instances (i.e. Figure 8). This drop in significance may appear surprising, but it is actually expected as the real patterns may or may not be meaningful depending on the context. On the other hand, we expect the fake patterns to be less likely to be meaningful. Therefore it is expected that there will be differences between the groups but less significant than Figure 8. This result might also be an indication to the fact that the question of "meaningfulness" is comparatively more subjective and vague than the previous scenario. Nevertheless, the statistical significance implies that the algorithm is more accurate than just random samples from the MoCap signal.

In the third research question, we ask if there is any improvement in the speaker's body language due to the use of our interface. In general, this is a difficult question to answer with strong affirmation. Typically, change of behavior is a gradual process and requires motivation and effort from within the person. Moreover, it may be hard to immediately notice any change in a person's behavior given a short exposure to a stimuli. Our interface could possibly make the speakers aware of their body language. However, how well they are able to internalize and reflect on the given insights is out of the current scope of the paper.

Nonetheless, to measure improvement in body language, we analyze how the speakers' self-ratings differ from the first speech to the last speech. The result is shown in Figure 10. The top plot of Figure 10 shows the average of self ratings in

[2]https://www.mturk.com/mturk/welcome

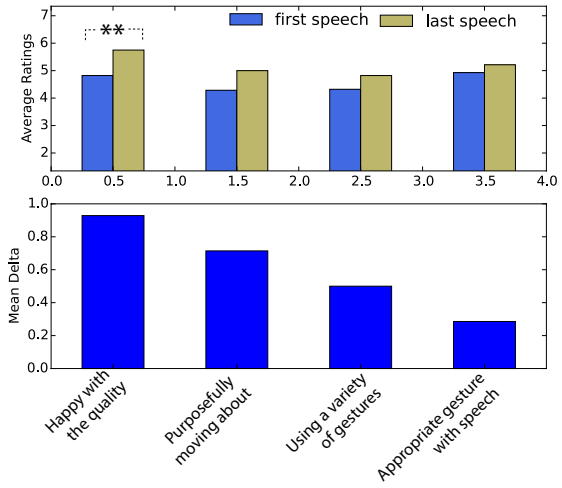

Figure 10: Participants' self ratings about their speech

the first video and the third video. Although there is an improvement in the average, only the first measure ("Overall, I am happy with the quality of my speech") is statistically significant ($p < 0.01$). We also computed the average *delta* as shown in the bottom plot of Figure 10. Mean delta is computed by taking away the ratings of first video from the rating of the third video for each participant, and then computing the average. The plot shows the mean delta is positive for all the self measures. These are evidence to the fact that there is an improvement over the participants' speaking. However, we cannot immediately claim that this improvement is a direct impact of using AutoManner. There are other considerations should be taken into account. Firstly, these are self-reports and the participants knew which one is their first speech and which one is last. That knowledge may lead to a bias for this result. Secondly, even if there are true improvements in their speaking, we never know if that is due to the usage of AutoManner or just due to practicing public speaking three times in the study.

Besides the measures as required for answering the research questions, we collect a few additional measures, as well. For example, we collect the participants' ratings on the accuracies of the patterns. We make some interesting observations while analyzing the responses of these measures as illustrated in Figure 11.

It is evident from this data that the participants could not actually differentiate between the fake patterns (placebo) and the real patterns (treatment). In addition, they sometimes rated patterns from the placebo interface higher than those from the treatment interface. This result is surprising especially given the fact that there is strong statistical significance for the difference in the mechanical turkers' annotations. As we shall discuss in the next subsection, this effect became more evident when we analyzed the participants' opinions in one-to-one interviews. In the Discussion section, we elaborate the reasons of this effect. We shall also describe how these two experiments focus on two different levels of the same question and thus one is inherently difficult to answer than another.

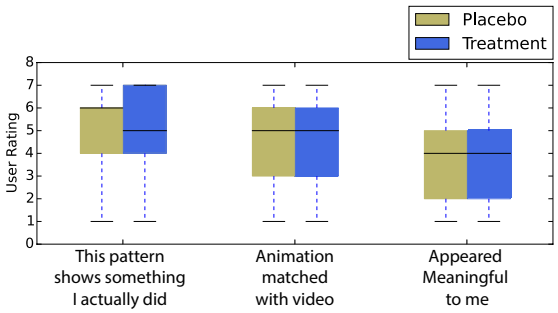

Figure 11: Participants' response to pattern-specific statements in the interface

SUBJECTIVE EVALUATION

We notice an overall trend of positive responses in the subjective interviews. When asked if the participants think they became aware of their body language, all of them replied in affirmation. One participant responded,

> "*It definitely made me more aware. It showed that I make the gestures sporadically. I was consistently showing the same gesture repeatedly throughout the three minutes of my talk.*"

While the participants highly praised the fact that they became aware of their body languages, they also raised questions about the accuracy of the interface. For example, one participant mentioned,

> "*Sometimes the skeleton and what I was doing didn't match ... and sometimes I wasn't sure whether or not, cuz there was part which did match and part which didn't match at all. So I'm not sure to say yes or not to it.*"

Some of the participants complained that the algorithm is actually picking the same patterns,

> "*It was a good idea here, but I'm not sure it picks up all of my gestures that well ... It seems like it's picking up all the same gestures. For example, I swayed a lot. It picked up the swaying but put it as five different gestures*"

However, another participant said, "*I would say, more than half are accurate.*". We thought these discrepancies were just due to the fact that the participants watched not only a treatment interface but a placebo interface as well. So perhaps they were referring to the errors in the placebo interface.

When we asked if the participants realized that there were two different styles in the interface—one real and another fake—none of them answered in affirmation. They appeared perplexed. One of the participant even argued that it was not possible because,

> "*I [the participant] remember both of them [the placebo and treatment interfaces] had the same thing [i.e. same pattern]!*".

Their answers to guess which interface showed the real pattern also appeared random. This provides strong evidence

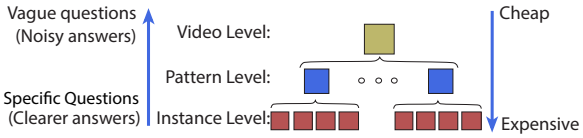

Figure 12: An illustration of our experience in mechanical turk data collection process

that the participants believed, in some cases, the placebo interface was as good as or even better than the treatment interface. We discuss more on this in the next section.

DISCUSSIONS

We became interested in knowing why the participants were unable to differentiate between the placebo and the treatment interfaces. In order to find the answer, we manually went through the interfaces. In this process, we found several patterns similar in both the placebo and the treatment interface. We found that the method that we used to generate the fake patterns was responsible for this problem. We generated the fake patterns by (uniformly) randomly picking a two-second window from the MoCap sequence and randomly assigning the time-instances for each pattern. We assumed, as the time-instances were completely random, it was a good strategy for designing the placebo interface. However, this argument is not valid when a single movement is repeated many times throughout the video. In that case, even a randomly selected window may pick up accurate gestures as the same pattern is repeated in the original sequence. In addition, the time instances may match in some cases due of the existence of high number of same patterns. Therefore, perhaps, our randomly sampled fake patterns, in some instances, might have overlapped with the real patterns. In retrospect, this outcome was not obvious during the initial stage of our exploration. We consider this as an important knowledge acquired in the process of experimentation to inform our future work.

A question still remains unanswered in the explanation above. Why the mechanical turk annotations were not as affected as the participants' annotations? The answer to this question resides in the way we collected data in these two experiments. Turkers annotated one time-instance at a time. When they looked at a certain time-instance, they only decided if the current movement in the video is matching with the skeletal movements or not. On the other hand, when the participants decide if a pattern matched with the video, they were presented with all the time-instances at once. In order to decide this match, they needed to watch several time-instances and summarize what they can *recall*. We hypothesize, these two processes are fundamentally different and thus it is unfair to make a direct comparison between the two results.

From this experience we gained interesting insights on designing behavior related questions in general as shown with an illustration in Figure 12. It is easier to answer specific questions that require less memory. On the other hand, it is difficult to answer questions where a person needs to decide an answer by remembering answers to a large quantity of smaller questions. For example, in our study, the task that asked the participant to match a particular time-instance of a video with the skeletal animation is simple; whereas, answering if a pattern is meaningful or not is context dependent and thus, hard. Unfortunately, it is more expensive to collect responses for more specific questions because they need a lot of questions to be answered. A well designed study needs to find a balance in this trade-off.

Finally, we sought answer to one last question: why did the participants like the interface and claim that they became aware of their body language if they were confused about differentiating the real and fake patterns? We argue that, although becoming self-aware is dependent on algorithmic accuracy, this dependence is not linearly related. The users only need a few correct samples to become aware of their body language.

FUTURE WORK

This work can help future experiments that follow a similar design. We think this experience has enough potential to generate numerous interesting questions related to study design and careful choice of questions in behavioral experiments. In the future, we shall continue our endeavor to design better interfaces with more appropriate choices of artificial intelligence components. We can possibly try other interesting unsupervised approaches for extracting body movement patterns. Unsupervised analysis of nonverbal behaviors is a lightly-explored and open field of research with numerous possibilities.

In this paper, we could not decisively determine whether the speakers actually improved their body language as a result of using this interface, or they just felt that they did better due to the interaction with a new technology. In the future, we shall try to evaluate this aspect using the opinions of public speaking experts. We shall also try to run a longitudinal study in order to tease out the novelty effect.

CONCLUSION

We presented an automated interface to make public speakers aware of their mannerisms. We used an unsupervised, *Shift Invariant Sparce Coding* (SISC) algorithm to automatically extract the common body movement patterns of the speaker. We showed these patterns as potential mannerism candidates. We designed a subtle question and answer-based scheme to provide non-judgmental feedback on mannerisms. Our experiments show that the speakers liked the interface, as they became aware of their body language. In addition, mechanical turk worker's annotations reveal that the algorithm we used in this system can extract human-interpretable patterns of body language.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their detailed and encouraging feedback. Thanks to Tergel Purevdorj for helpful comments on the interface. This work was supported in part by Grant W911NF-15-1-0542 with the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO). Special Acknowledgment to Microsoft Azure for providing the computational platform.

REFERENCES

1. Aggarwal, J. K., et al. Human activity analysis: A review. *ACM Computing Surveys (CSUR)* (2011).
2. Anderson, K., André, E., Baur, T., Bernardini, S., Chollet, M., Chryssafidou, E., Damian, I., Ennis, C., Egges, A., Gebhard, P., et al. The tardis framework: intelligent virtual agents for social coaching in job interviews. In *Advances in Computer Entertainment*. Springer, 2013, 476–491.
3. Batrinca, L., Stratou, G., Shapiro, A., Morency, L.-P., and Scherer, S. Cicero-towards a multimodal virtual audience platform for public speaking training. In *Intelligent Virtual Agents*, Springer (2013), 116–128.
4. Battiti, R. Accelerated backpropagation learning: Two optimization methods. *Complex systems 3*, 4 (1989), 331–342.
5. Beck, A., and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences 2*, 1 (2009), 183–202.
6. Chen, L., Feng, G., Joe, J., Leong, C. W., Kitchen, C., and Lee, C. M. Towards automated assessment of public speaking skills using multimodal cues. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ACM (2014), 200–203.
7. Cheng, G., et al. Advances in human action recognition: A survey. *arXiv preprint arXiv:1501.05964* (2015).
8. Chollet, M., Wörtwein, T., Morency, L.-P., Shapiro, A., and Scherer, S. Exploring feedback strategies to improve public speaking: an interactive virtual audience framework. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM (2015), 1143–1154.
9. D'Arcy, J. *Technically speaking: A guide for communicating complex information*. Battelle Press Columbus, OH, 1998.
10. de Gelder, B. Why bodies? twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences 364*, 1535 (2009), 3475–3484.
11. DiMatteo, M. R., Hays, R. D., and Prince, L. M. Relationship of physicians' nonverbal communication skill to patient satisfaction, appointment noncompliance, and physician workload. *Health Psychology 5*, 6 (1986), 581.
12. Fay, M. P., and Proschan, M. A. Wilcoxon-mann-whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics surveys 4* (2010), 1.
13. Hoogterp, B. *Your Perfect Presentation: Speak in Front of Any Audience Anytime Anywhere and Never Be Nervous Again*. McGraw-Hill Education, 2014.
14. Hoque, M. E., Courgeon, M., Martin, J.-C., Mutlu, B., and Picard, R. W. Mach: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, ACM (2013), 697–706.
15. Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology 24*, 6 (1933), 417.
16. Knapp, M., Hall, J., and Horgan, T. *Nonverbal communication in human interaction*. Cengage Learning, 2013.
17. Likert, R. A technique for the measurement of attitudes. *Archives of psychology* (1932).
18. Lucas, S. E. *The art of public speaking*. International Book Publishing Company, 2008.
19. Metaxas, D., and Zhang, S. A review of motion analysis methods for human nonverbal communication computing. *Image and Vision Computing* (2013).
20. Mitra, T., Hutto, C., and Gilbert, E. Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM (2015), 1345–1354.
21. Mørup, M., et al. Shift invariant sparse coding of image and music data. Tech. Rep. IMM2008-04659, Technical University of Denmark, 2008.
22. Murphy, J. *The power of your subconscious mind*. Courier Corporation, 2012.
23. Naim, I., Tanveer, M. I., Gildea, D., and Hoque, M. E. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. *Automatic Face and Gesture Recognition (FG)* (2015).
24. Nguyen, A.-T., Chen, W., and Rauterberg, M. Online feedback system for public speakers. In *E-Learning, E-Management and E-Services (IS3e), 2012 IEEE Symposium on*, IEEE (2012), 1–5.
25. Niebles, J. C., Wang, H., and Fei-Fei, L. Unsupervised learning of human action categories using spatial-temporal words. *International journal of computer vision 79*, 3 (2008), 299–318.
26. Park, S., Shoemark, P., and Morency, L.-P. Toward crowdsourcing micro-level behavior annotations: the challenges of interface, training, and generalization. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, ACM (2014), 37–46.
27. Pfister, T., and Robinson, P. Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis. *Affective Computing, IEEE Transactions on 2*, 2 (2011), 66–78.
28. Ranganath, R., Jurafsky, D., and McFarland, D. It's not you, it's me: detecting flirting and its misperception in speed-dates. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, Association for Computational Linguistics (2009), 334–342.

29. Schreiber, L. M., Paul, G. D., and Shibley, L. R. The development and test of the public speaking competence rubric. *Communication Education 61*, 3 (2012), 205–233.

30. Shim, H. S., Park, S., Chatterjee, M., Scherer, S., Sagae, K., and Morency, L.-P. Acoustic and para-verbal indicators of persuasiveness in social multimedia. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, IEEE (2015), 2239–2243.

31. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., and Moore, R. Real-time human pose recognition in parts from single depth images. *Communications of the ACM 56*, 1 (2013), 116–124.

32. Strangert, E., and Gustafson, J. What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations. In *INTERSPEECH*, vol. 8 (2008), 1688–1691.

33. Tanaka, H., Sakti, S., Neubig, G., Toda, T., Negoro, H., Iwasaka, H., and Nakamura, S. Automated social skills trainer. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, ACM (2015), 17–27.

34. Tanveer, M. I., Lin, E., and Hoque, M. E. Rhema: A real-time in-situ intelligent interface to help people with public speaking. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, ACM (2015), 286–295.

35. Tanveer, M. I., Liu, J., and Hoque, M. E. Unsupervised extraction of human-interpretable nonverbal behavioral cues in a public speaking scenario. In *ACM Multimedia (ACMMM'15)* (2015).

36. Toastmasters International. Gestures: Your body speaks. Online Document. Available at `http://web.mst.edu/~toast/docs/Gestures.pdf`, 2011.

37. Vinciarelli, A., Pantic, M., and Bourlard, H. Social signal processing: Survey of an emerging domain. *Image and Vision Computing 27*, 12 (2009), 1743–1759.

38. Wilson, T. D. *Strangers to ourselves*. Harvard University Press, 2004.

39. Zhang, Z. Microsoft kinect sensor and its effect. *MultiMedia, IEEE 19*, 2 (2012), 4–10.

40. Zhou, F., et al. Aligned cluster analysis for temporal segmentation of human motion. In *FG'08* (2008).