Automated Prediction and Analysis of Job Interview Performance: The Role of What You Say and How You Say It

Iftekhar Naim[1], M. Iftekhar Tanveer[2], Daniel Gildea[1], and Mohammed (Ehsan) Hoque[1,2]
[1] ROC HCI, Department of Computer Science, University of Rochester
[2] ROC HCI, Department of Electrical and Computer Engineering, University of Rochester

***Abstract*— Ever wondered why you have been rejected from a job despite being a qualified candidate? What went wrong? In this paper, we provide a computational framework to quantify human behavior in the context of job interviews. We build a model by analyzing 138 recorded interview videos (total duration of 10.5 hours) of 69 internship-seeking students from Massachusetts Institute of Technology (MIT) as they spoke with professional career counselors. Our automated analysis includes facial expressions (e.g., smiles, head gestures), language (e.g., word counts, topic modeling), and prosodic information (e.g., pitch, intonation, pauses) of the interviewees. We derive the ground truth labels by averaging over the ratings of 9 independent judges. Our framework automatically predicts the ratings for interview traits such as excitement, friendliness, and engagement with correlation coefficients of 0.73 or higher, and quantifies the relative importance of prosody, language, and facial expressions. According to our framework, it is recommended to speak more fluently, use less filler words, speak as "we" (vs. "I"), use more unique words, and smile more.**

I. Introduction

Imagine the following scenario in which two students, John and Matt, were individually asked to discuss their leadership skills in a job interview. John responded with the following:

> "*One semester ago, I was part of a team of ten students [stated in a loud and clear voice]. We worked together to build an autonomous playing robot. I led the team by showing how to program the robot. The students did a wonderful job [conveyed excitement with tone]! In ten weeks, we made the robot play soccer. It was a lot of fun. [concluded with a smile]".*

Matt responded with the following:

> "*Umm ... [paused for 2 seconds] last semester I led a group in a class project on robot programming. It was a totally crazy experience. The students almost did nothing until the last moment. ... Umm ... Basically, I had to intervene at that point and led them to work hard. Eventually, this project was completed successfully. [looking away from the interviewer]*".

Who do you think received higher ratings?

Most would agree that the first interviewee, John, provided more enthusiastic and engaging answer. We can easily interpret the meaning of our verbal and nonverbal behavior during face-to-face interactions. However, we often can not quantify how the combination of these behaviors affect our interpersonal communications. Over many years, social psychologists and career counselors have accumulated knowledge and guidelines on succeeding in interviews [1]. For example, research has shown that smiling, a purposeful, confident voice, and comfortable eye contact contribute positively to our interpersonal communications. These guidelines are largely based on intuition, experience, and studies involving laborious manual encoding of nonverbal behaviors on a limited amount of data [1]. Despite the significant effort, automated and objective quantification of our social behavior remains a challenging problem.

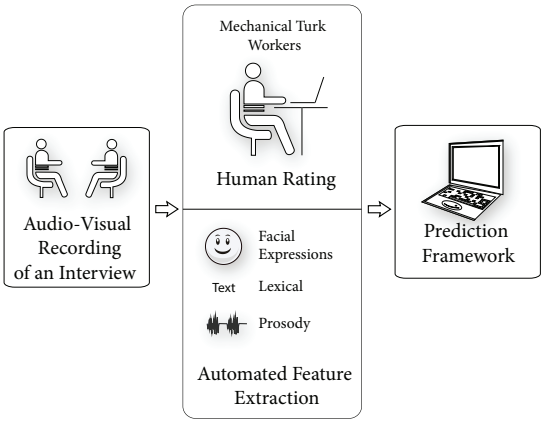

Fig. 1. Framework of Analysis. Human raters labeled interviewee performance by watching videos of job interviews. A total of 82 features were extracted from those videos. A framework was built to predict performance ratings and to gain insight into the characteristics of a good interviewee.

A common perception surrounding job interviews is that the content of the interviewee's answers is the most important determinant for success. However, empirical studies show that nonverbal behavior is as important as the verbal response in job interviews [1], [2]. Nonverbal behaviors are subtle, fleeting, subjective, and sometimes even contradictory, posing a significant challenge for any prediction framework. Even a simple facial expression such as a smile can elicit different meanings, such as delight, rapport, sarcasm, and even frustration [3]. The style of speaking, prosody, and language reflect valuable information about one's personality and mental state [4]. Understanding the relative influence of these individual modalities can provide crucial insight regarding job interviews.

In this paper, we attempt to answer the following research questions by analyzing audio-visual recordings of 138 inter-

view sessions with 69 individuals:

- Can we automatically quantify verbal and nonverbal behavior, and assess their role in the overall rating of job interviews?
- Can we build a computational framework that can automatically predict the overall rating of a job interview given the audio-visual recordings?
- Can we infer the relative importance of language, facial expressions, and prosody (intonation)?
- Can we make automated recommendations on improving social traits such as excitement, friendliness, and engagement in the context of a job interview?

To answer these research questions, we designed and implemented an automated prediction framework for quantifying the outcome of job interviews, given the audio-visual recordings. The proposed prediction framework (Figure 1) automatically extracts a diverse set of multimodal features (lexical, facial, and prosodic), and quantifies the overall interview performance, the likelihood of getting hired, and 14 other social traits relevant to the interview process. Our system is capable of predicting the overall rating of a job interview with a correlation coefficient $r = 0.70$ and AUC = 0.81 (baseline 0.50) when tested on a dataset of 138 interview videos of 69 participants. We can also predict different social traits such as engagement, excitement, and friendliness with even higher accuracy ($r \geq 0.73$, AUC > 0.80). We investigate the relative weights of the individual verbal and non-verbal features, and quantify the relative importance of language, prosody, and facial expressions. According to our analysis, prosody plays the most important role in the context of job interviews.

II. Background Research

A. Nonverbal Behavior Recognition

There has been a significant amount of work for automatic recognition of nonverbal behavioral patterns and social cues. Given the challenges of data collection and multimodal data analysis, most of the existing work focuses on a single behavioral modality, such as prosody [5], [6], facial expression [7], gesture [8], and word usage pattern [9]. Ranganath et al. [10] proposed a framework that predicts personality traits such as awkwardness, assertiveness, flirtatiousness, and friendliness using a combination of prosodic and lexical features. Similarly, Kapoor et al. [4] and Pianesi et al. [11] proposed systems to recognize different social and personality traits by exploiting only prosodic and visual features.

Sanchez et al. [12] proposed a system for predicting eleven different social moods (e.g., surprise, anger, happiness) from YouTube video monologues, which consist of different social dynamics than in face-to-face interactions. The most relevant work is the one by Nguyen et al. [13], who proposed a computational framework to predict the hiring decision using non-verbal behavioral cues extracted from a dataset of 62 interview videos. Our work extends the current state-of-the-art and generates new knowledge by incorporating three different modalities (prosody, language, and facial expressions), and sixteen different social traits (e.g., friendliness, excitement, engagement), and quantifies the interplay and relative influences of these different modalities for each of the different social traits. Furthermore, by analyzing the relative feature weights learned by our regression models, we obtain valuable insights about behaviors that are recommended for succeeding in job interviews (Section V-B.3).

Fig. 2. The experimental setup for collecting audio-visual recordings of the mock interviews. Camera #1 recorded the video and audio of the interviewee, while Camera #2 recorded the interviewer.

B. Social Coaching for Job Interviews

Several systems have been proposed for training the necessary social skills to succeed in job interviews [14], [15], [16]. Hoque et al. [14] developed MACH (My Automated Conversation coacH), which allows users to improve social skills by interacting with a virtual agent. Anderson et al. [15] proposed an interview coaching system, TARDIS, which presents the training interactions as a scenario-based "serious game". The TARDIS framework incorporates a submodule named NovA (NonVerbal behavior Analyzer) [16] that can recognize several lower level social cues: *hands-to-face*, *looking away*, *postures*, *leaning forward/backward*, *gesticulation*, *voice activity*, *smiles*, and *laughter*. Using videos that are manually annotated with these ground truth social cues, NovA trains a Bayesian Network that can infer higher-level mental traits (e.g., stressed, focused, engaged). Automated prediction of higher-level traits using social cues remains part of their future work.

Our framework extends the state-of-the-art by (1) quantifying the relative influences of different low-level features on the interview outcome, (2) learning regression models to predict interview ratings and the likelihood of hiring using automatically extracted features, and (3) predicting several other high-level personality traits such as engagement, friendliness, and excitement.

III. MIT Interview Dataset

We used the *MIT Interview Dataset* [14], which consists of 138 audio-visual recordings of mock interviews with internship-seeking students from Massachusetts Institute of Technology (MIT).

A. Data Collection

1) Study Setup: The mock interviews were conducted in a room equipped with a desk, two chairs, and two wall-mounted cameras (Figure 2). The two cameras with

microphones were used to capture the facial expressions and the audio conversations during the interview.

2) Participants: Initially 90 juniors participated in the mock interviews. All participants were native English speakers. The interviews were conducted by two professional career counselors who had over five years of experience. For each participant, two rounds of mock interviews were conducted: before and after interview intervention. Each individual received $50 for participating. Furthermore, as an incentive for the participants, we promised to forward the resume of the top 5% candidates to several sponsor organizations (Deloitte, IDEO, and Intuit) for considerations for summer internships. After the data collection, 69 (26 male, 43 female) of the 90 initial participants permitted the use of their video recordings for research purposes and to be shared with other researchers.

3) Procedure: During each interview session, the counselor asked each interviewee five questions. No job description was given to the interviewees. The five questions were chosen to assess behavioral and social skills only. The total duration of our interview videos is nearly 10.5 hours (on average, 4.7 minutes per interview, for 138 interview videos). To our knowledge, this is the largest collection of job interview videos conducted by professional counselors under realistic settings.

B. Data Labeling

The subjective nature of human judgment makes it difficult to collect ground truth for interview ratings. Due to the nature of the experiment, the counselors interacted with each interviewee twice—before and after the intervention, and provided feedback after each session. The process of feedback and the way the interviewees responded to the feedback may have had an influence on the counselor's ratings. In order to remove the bias introduced by the interaction, we used Amazon Mechanical Turk workers to rate the interview performance. Each Turker watched the interview videos and rated the performances of the interviewees by answering 16 assessment questions (Figure 3) on a seven point Likert scale[1]. The questions about "Overall Rating" and "Recommend Hiring" represent the overall performance. The remaining questions have been selected to evaluate several high-level behavioral dimensions such as warmth (e.g., "friendliness", "smiling"), presence (e.g., "engagement", "excitement", "focused"), competence (e.g. speaking rate), and content (e.g., "structured"). Apart from being more objective, the Mechanical Turk workers could pause and replay the video, allowing them to rate more thoroughly. However, the Turkers' ratings are more likely to be similar to the "audience" ratings, as opposed to being the "expert ratings".

We first selected 10 Turkers out of 25, based on how well they agreed with the career counselors on the five control videos. Out of these 10 selected Turkers, one did not finish all the rating tasks, leaving us with 9 ratings per video. We have automatically estimated the quality of individual workers using an EM-style optimization algorithm (described in the Appendix), and estimated a weighted average of their scores as the ground truth ratings.

[1]Appendix Table I describes the 16 assessment questions.

IV. Prediction Framework

For the prediction framework, we automatically extracted 82 features from the interview videos, and trained two regression models: Support Vector Regression (SVR) [17] and Lasso [18]. The objective of this training is twofold: first, to predict the Turkers' ratings on the overall performance and each behavioral trait, and second, to quantify and gain meaningful insights on the relative importance of individual features for each trait.

A. Feature Extraction

We collected three types of features for each interview video: (1) prosodic features, (2) lexical features, and (3) facial features. We selected features that have been shown to be relevant for job interviews [1], and other social interactions [12], [10], [19]. For extracting reliable lexical features, we chose not to use automated speech recognition. Instead, we transcribed the videos by hiring Amazon Mechanical Turk workers, who were specifically instructed to include filler and disfluency words such as "uh", "umm", and "like", in the transcription. We also collected a wide range of prosodic and facial features.

1) Prosodic Features: Prosody reflects our speaking style, particularly the rhythm and the intonation of speech. Prosodic features have been shown to be effective for social intent modeling [5], [6]. We extracted prosodic features of the interviewee's speech using the open-source speech analysis tool PRAAT [20]. Each prosodic feature is first collected over the duration corresponding to each individual answer by the interviewee, and then averaged over her/his five answers. While averaging the prosodic features over all the answers reduces the dimensionality of the feature space, it also loses the temporal structure in prosody.

The important prosodic features include the pitch information, vocal intensities, characteristics of the first three formants, and spectral energy, which have been reported to reflect our social traits [6]. For reflecting the vocal pitch, we extracted the mean and the standard deviation (SD) of fundamental frequency F0 (F0 MEAN and F0 SD), the minimum and maximum values (F0 MIN, F0 MAX), and the total range (F0 MAX - F0 MIN). We extracted similar features for voice intensity and the first 3 formants. Additionally, we collect several other prosodic features such as pause duration, percentage of unvoiced frames, jitter (irregularities in pitch), shimmer (irregularities in vocal intensity), and percentage of breaks in speech.

2) Lexical features: Lexical features provide valuable information regarding interview content and speaking style. One of the most commonly used lexical features is the counts of individual words. However, incorporating word counts often results in sparse high-dimensional feature vectors,

and suffers from the "curse of dimensionality" problem, especially for a limited sized corpus.

We address this challenge with two techniques. First, instead of using raw unigram counts, we employed counts of various psycholinguistic word categories defined by the tool "Linguistic Inquiry Word Count" (LIWC) [21]. The LIWC categories include negative emotion terms (e.g., sad, angry), positive emotion terms (e.g., happy, kind), different function word categories (e.g., articles, quantifiers), pronoun categories (e.g., I, we, they), and various content word categories (e.g., anxiety, insight). We selected 23 such LIWC word categories, which is significantly smaller than the number of individual words.

Although the hand coded LIWC lexicon has proven to be useful for modeling many different social behaviors [10], the lexicon is predefined and may not cover many important aspects of job interviews. To address this challenge, we applied Latent Dirichlet Allocation (LDA) [22] to automatically learn common topics from our interview dataset. We set the number of topics to 20. For each interview, we estimate the relative weights of these learned topics, and use these weights as lexical features.

Finally, we collected additional features related to our linguistic and speaking skills, such as: *wpsec* (words per second), *upsec* (unique words per second), *wc* (word count), *uc* (unique word count), and *fpsec* (filler words per second). Similar speaking rate and fluency features were exploited by Zechner et al. [19] in the context of automated scoring of non-native speech in TOEFL practice tests.

3) Facial features: We extracted facial features for the interviewees from each video frame. First, faces were detected using the Shore [23] framework. We trained an AdaBoost classifier to distinguish between the neutral and smiling faces. The classifier output is normalized in the range [0,100], where 0 represents no smile, and 100 represents full smile. We took an average over the smile intensities from individual frames, and use this as a feature. We also extracted head gestures such as nods and shakes [14] from each video frame, and treated their average values as features.

4) Feature Normalization: We concatenate the three types of features described above, and obtain one combined feature vector. To remove any possible bias related to the range of values associated with a feature, we normalized each feature to have zero mean and unit variance.

B. Score Prediction from Extracted Features

Using the features described in the previous section, we train regression models to predict the overall interview scores and other interview-specific traits (e.g., excitement, friendliness, engagement, and awkwardness). We experimented with many different regression models: SVR, Lasso, L_1 Regularized Logistic Regression, and Gaussian Process. We will only discuss SVR (with linear kernel) and Lasso, which achieved the best results with our dataset.

V. Results

First, we analyze the quality and reliability of Turkers' ratings by observing how well the Turkers agree with each other

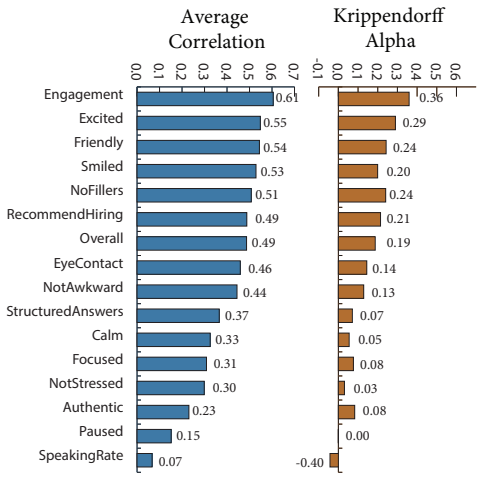

Fig. 3. The inter-rater agreement among the Turkers, measured by the Krippendorff's Alpha (varies in the range $[-1.0, 1.0]$) and the average one-vs-rest correlation of their ratings (range $[-1.0, 1.0]$).

(Section V-A). Next, we present the prediction accuracies for the trained regression models (SVR and Lasso) based on automatically extracted features (Section V-B). Finally, we analyze the weights of the features in our trained models, and quantify the relative importance of the behavioral features to overall interview performance (Section V-B.2).

A. Inter-Rater Agreement

To assess the quality of the ratings, we calculate Krippendorff's Alpha [24] for each trait. In this case, Krippendorff's Alpha is more meaningful than the frequently used Fleiss' Kappa [25], as the ratings are ordinal values (on a 7-point Likert scale). The value of Krippendorff's Alpha can be any real number in the range $[-1.0, 1.0]$, with 1.0 being the perfect agreement and -1.0 being absolute disagreement among the raters. We also estimate the correlation of each Turker's rating with the mean rating by the other Turkers for each trait. Figure 3 shows that some traits have relatively good inter-rater agreement among the Turkers (e.g., "engagement", "excitement", "friendliness"). Some other traits such as: "stress", "authenticity", "speaking rate", and "pauses" have low inter-rater agreement. It may be difficult for the Turkers to agree on the subjective interpretation of those attributes without interacting with the participant.

B. Prediction using Automated Features

1) Prediction Accuracy using Trained Models: Given the feature vectors associated with each interview video, we trained 16 regression models for predicting the ratings for the 16 traits or rating categories. The entire dataset has a total of 138 interview videos (for the 69 participants, 2 interviews for each participant). To avoid any artifacts related to how we split the data into training and test sets, we performed 1000 independent trials. In each trial, we randomly chose 80% of the videos for training, and the remaining 20% for testing. We report our results averaged over these 1000 independent trials. For each of the traits, we used the same set of features, and the model automatically learned the feature weights.

TABLE I

THE AVERAGE AREA UNDER THE ROC CURVE.

Trait	SVR	LASSO
Excited	0.91	0.88
Engagement	0.87	0.86
Smiled	0.85	0.85
Recommend Hiring	0.82	0.79
No Fillers	0.82	0.86
Overall	0.81	0.78
Structured Answers	0.81	0.80
Friendly	0.80	0.79
Focused	0.80	0.68
Not Awkward	0.79	0.77
Paused	0.75	0.74
Eye Contact	0.69	0.61
Authentic	0.69	0.64
Calm	0.68	0.65
Speaking Rate	0.63	0.54
Not Stressed	0.63	0.58

We measure prediction accuracy by the correlation coefficients between the true ratings (weighted average of Turkers' ratings) and the predicted ratings on the test sets. Figure 4 displays the correlation coefficients for different traits, both with SVR and Lasso. The traits are shown in the order of their correlation coefficients obtained by SVR. We performed well in predicting overall performance and hiring recommendation scores ($r = 0.70$), which are the two most important scores for interview decision. Furthermore, we can predict traits such as engagement, excitement, and friendliness with 0.73 or higher correlation coefficients. We would like to point out that the interview questions asked in our training dataset are chosen to be independent of any job specifications or skill requirements. Therefore, the ratings predicted by our model are based on social and behavioral skills only, and they may differ from a hiring manager's opinion, given specific job requirements.

We also evaluated the trained regression models for a binary classification task by splitting the interviews into two classes by the median rating for each trait. Any interview with a score higher than the median value for a particular trait is considered to be in the positive class (for that trait), and the rest are placed in the negative class. We estimate the area

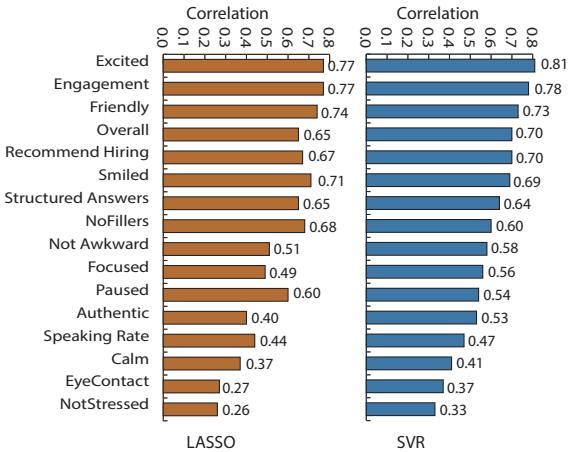

Fig. 4. Regression coefficients using two different methods: Support Vector Regression (SVR) and Lasso.

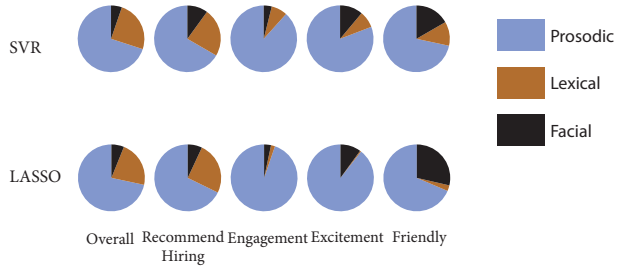

Fig. 5. Relative proportion of the top twenty prosodic, lexical, and facial features as learned by SVR and Lasso.

under the Receiver Operator Curve (ROC) by varying the discrimination threshold (Table I). The baseline area under the curve (AUC) value is 0.50, as we split the classes by the median value. Again, we observed high accuracies for engagement, excitement, friendliness, hiring recommendation, and the overall score (AUC > 0.80).

When we examine the traits with lower prediction accuracy, we observe: (1) either we have low interrater agreement for these traits, which indicates unreliable ground truth data (e.g., calm, stressed, speaking rate, pause), or (2) we lack key features necessary to predict these traits (e.g., eye contact). In the absence of eye tracking information (which is very difficult to obtain automatically from videos), we do not have enough informative features to predict eye contact.

2) Feature Analysis: We examined the relative weights of individual features in our regression models, and obtained valuable insights on the essential constituents of a job interview. We considered five traits with relatively high prediction accuracy ($AUC > 0.80$, $r > 0.70$) : overall score, hiring recommendation, excitement, engagement, and friendliness. For each of these five traits, we examined the top twenty features in the order of descending weight magnitudes, and estimated the summation of the weight magnitudes of the features in each of the three categories: prosodic, lexical, and facial features. The relative proportions (Figure 5) show that both SVR and Lasso assign higher weights to prosodic features while predicting engagement and excitement, which matches our intuition that excitement and engagement are usually expressed by our intonation. For both models, the relative weights of features for predicting the "overall rating" and "recommend hiring" are similar, which is expected, as these two traits are highly correlated.

Since we had only three facial features (smile, nod, and shake), the relative weights for facial features are much lower. However, facial features, particularly the smile, were found to be significant for predicting friendliness.

3) Recommendation from our Framework: To better understand the recommended behavior in job interviews, we analyze the feature weights in our regression model. Positive weights with higher magnitudes can potentially indicate elements of a successful job interview. The negative weights indicate behaviors we should avoid.

We sort the features by the magnitude of their weights and examine the top twenty features, excluding the topic features (Appendix Table IV and V). We found that people having higher speaking rate (higher words per second (*wpsec*), total

number of words (*wc*), and total number of unique words (*uc*)) are perceived as better candidates. People who speak more fluently and use fewer filler words (lower number of filler words per second (*fpsec*), total number of filler words (*Fillers*), total number non-fluency words (*Non-fluencies*), less unvoiced region in speech (*%Unvoiced*), and fewer breaks in speech (*%Breaks*)) are perceived as better candidates. We also find that higher interview score correlates with higher usage of words in LIWC category "*They*" and "*We*", and lower usage of words related to "*I*". The overall interview performance and likelihood of hiring correlate positively with the proportion of positive emotion words, and negatively with the proportion of negative emotion words, which agrees with our experience. Individuals who smiled more were rated higher in job interviews. Finally, those speaking with a higher proportion of quantifiers (e.g., best, every, all, few), perceptual words (e.g. see, observe, know), and other functional word classes (articles, prepositions, conjunctions) obtained higher scores. As we've seen earlier, features related to prosody and speaking style are more important for excitement and engagement. Particularly the amplitude and range of the voice intensity and pitch had high positive weights in our prediction model. Finally, besides smiling, people who spoke more words associated with *"We"* than *"I"* were perceived as being friendlier.

VI. Conclusion and Future Work

We present an automated prediction framework for quantifying social skills in job interviews. The proposed models show encouraging results for predicting human interview ratings and several mental traits such as engagement, excitement, and friendliness. Furthermore, we extracted quantitative knowledge about recommended behaviors in job interviews, which is consistent with past literature and agrees with our intuition.

One of our immediate next steps will be to integrate the proposed prediction module with existing automated conversational systems such as MACH [14] to allow feedback to the users. With the knowledge presented in this paper, we could train a system to help underprivileged youth receive feedback on job interviews that require a significant amount of social skills. The framework could also be expanded to help people with social difficulties, train customer service professionals, or even help medical professionals with telemedicine.

While limiting the study to undergraduate students from a particular institute helped control for any possible variability, it might have introduced a selection bias in our dataset. In future, we would like to conduct a more comprehensive study over a diverse population group to address this limitation.

The outcome of job interviews often depends on a subtle understanding of the interviewee's response. In our dataset, we noticed interviews in which a momentary mistake (e.g., the use of a swear word) ruined the interview outcome. Due to the rare occurrences of such events, it is difficult to model these phenomena, and perhaps anomaly detection techniques could be more effective instead. Extending our prediction framework for quantifying these diverse and complex cues can provide valuable insight and understanding regarding job interviews and human behavior in general.

References

[1] A. I. Huffcutt, J. M. Conway, P. L. Roth, and N. J. Stone, "Identification and meta-analytic assessment of psychological constructs measured in employment interviews." *Journal of Applied Psychology*, vol. 86, no. 5, p. 897, 2001.

[2] A. Mehrabian, *Silent messages*. Belmont: Wadsworth, 1971.

[3] M. E. Hoque, D. J. McDuff, and R. W. Picard, "Exploring temporal patterns in classifying frustrated and delighted smiles," *Affective Computing, IEEE Transactions on*, vol. 3, no. 3, pp. 323–334, 2012.

[4] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *ACM Multimedia*, 2005, pp. 677–682.

[5] V. Soman and A. Madan, "Social signaling: Predicting the outcome of job interviews from vocal tone and prosody," in *ICASSP*. Dallas, Texas, USA: IEEE, 2010.

[6] R. W. Frick, "Communicating emotion: The role of prosodic features." *Psychological Bulletin*, vol. 97, no. 3, p. 412, 1985.

[7] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3d facial expression recognition: A comprehensive survey," *Image and Vision Computing*, vol. 30, no. 10, pp. 683–697, 2012.

[8] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *Affective computing and intelligent interaction*. Springer, 2007, pp. 71–82.

[9] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[10] R. Ranganath, D. Jurafsky, and D. A. McFarland, "Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates," *Computer Speech & Language*, vol. 27, no. 1, pp. 89–115, 2013.

[11] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal recognition of personality traits in social interactions," in *ICMI*. ACM, 2008, pp. 53–60.

[12] D. Sanchez-Cortes, J.-I. Biel, S. Kumano, J. Yamato, K. Otsuka, and D. Gatica-Perez, "Inferring mood in ubiquitous conversational video," in *MUM*. New York, NY, USA: ACM, 2013, pp. 22:1–22:9.

[13] L. Nguyen, D. Frauendorfer, M. Mast, and D. Gatica-Perez, "Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior," *Multimedia, IEEE Transactions on*, vol. 16, no. 4, pp. 1018–1031, June 2014.

[14] M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard, "MACH: My automated conversation coach," in *Ubicomp*. ACM, 2013, pp. 697–706.

[15] K. Anderson, E. André, T. Baur *et al.*, "The TARDIS framework: intelligent virtual agents for social coaching in job interviews," in *Advances in Computer Entertainment*. Springer, 2013, pp. 476–491.

[16] T. Baur, I. Damian, F. Lingenfelser, J. Wagner, and E. André, "Nova: Automated analysis of nonverbal signals in social interactions," in *Human Behavior Understanding*. Springer, 2013, pp. 160–171.

[17] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[18] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[19] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, vol. 51, no. 10, pp. 883 – 895, 2009.

[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (Version 5.4.04)[Computer Program]. retrieved Dec 28, 2014," 2014.

[21] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, "The development and psychometric properties of LIWC2007," *Austin, TX, LIWC. Net*, 2007.

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[23] B. Froba and A. Ernst, "Face detection with the modified census transform," in *Automatic Face and Gesture Recognition (FG)*. IEEE, 2004, pp. 91–96.

[24] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educational and Psychological Measurement*, vol. 30, no. 1, pp. 61–70, 1970.

[25] J. L. Fleiss, B. Levin, and M. C. Paik, "The measurement of interrater agreement," *Statistical methods for rates and proportions*, vol. 2, pp. 212–236, 1981.

Appendix for the Paper: "Automated Prediction and Analysis of Job Interview Performance: The Role of What You Say and How You Say It"

Iftekhar Naim[1], M. Iftekhar Tanveer[2], Daniel Gildea[1], and Mohammed (Ehsan) Hoque[1,2]
[1] ROC HCI, Department of Computer Science, University of Rochester
[2] ROC HCI, Department of Electrical and Computer Engineering, University of Rochester

Appendix
Estimating Turker Reliability

We aim to automatically estimate the reliability of each Turker, and the ground truth ratings based on the Turkers' ratings. We adapt a simplified version of the existing latent variable model by Raykar et al. [1], that treats the reliability of each Turker and the ground truth ratings as latent variables, and estimate their values using an EM-style iterative optimization technique.

Let $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ be a dataset containing N feature vectors $\mathbf{x}_i$ (one for each interview video), for which the ground truth label y_i is unknown. We acquire subjective labels $\{y_i^1, \ldots, y_i^K\}$ from K Turkers on a seven point likert scale, i.e., $y_i^j \in \{1, 2, \ldots, 7\}$. Given this dataset $\mathcal{D}$, our goal is to learn the true rating (y_i) and also the reliability of each worker (λ_j).

To simplify the estimation problem, we assume the Turkers' ratings as real numbers, i.e., $y_i^j \in \mathbb{R}$. We also assume that each Turker's rating is a noisy version of the true rating $y_i \in \mathbb{R}$, perturbed via additive Gaussian noise. Therefore, the probability distribution for the y_i^j:

$$Pr[y_i^j | y_i, \lambda_j] = \mathcal{N}(y_i^j | y_i, 1/\lambda_j) \tag{1}$$

where λ_j is the unknown inverse-variance and the measure of reliability for the j^{th} Turker. By taking the logarithm on both sides and ignoring constant terms, we get the log-likelihood function:

$$L = \sum_{i=1}^{N} \sum_{j=1}^{K} \left[\frac{1}{2} \log \lambda_j - \frac{\lambda_j}{2} (y_i^j - y_i)^2 \right] \tag{2}$$

The log-likelihood function is non-convex in y_i and λ_j variables. However, if we fix y_i, the log-likelihood function becomes convex with respect to λ_j, and vice-versa. Assuming λ_j fixed, and setting $\frac{\partial L}{\partial y_i} = 0$, we obtain the update rule:

$$y_i = \frac{\sum_{j=1}^{K} \lambda_j y_i^j}{\sum_{j=1}^{K} \lambda_j} \tag{3}$$

Similarly, assuming y_i fixed, and setting $\frac{\partial L}{\partial \lambda_j} = 0$, we obtain the update rule:

$$\lambda_j = \frac{\sum_{i=1}^{N} (y_i^j - y_i)^2}{N} \tag{4}$$

We alternately apply the two update rules for y_i and λ_j for $i = 1, \ldots, N$ and $j = 1, \ldots, K$ until convergence. After convergence, the estimated y_i values are treated as ground truth ratings and used for training our prediction models.

Appendix
List of Questions Asked to Interviewees

During each interview session, the counselor asked an interviewee the following five questions in the following order:

Q1. So please tell me about yourself.
Q2. Tell me about a time when you demonstrated leadership.
Q3. Tell me about a time when you were working with a team and faced a challenge. How did you overcome the problem?
Q4. What is one of your weaknesses and how do you plan to overcome it?
Q5. Now, why do you think we should hire you?

Appendix
List of Assessment Questions Asked to Mechanical Turk Workers

Each Mechanical Turk worker was asked 16 questions to assess the performance of the interviewee. The list of these 16 questions is presented in Table I.

Appendix
List of Prosodic and Lexical Features

In this section, we present a list of all the prosodic and lexical features used in our framework. Table II lists all the prosodic features used in our framework. Table III presents all the LIWC lexical features.

Appendix
Overview of Support Vector Regression (SVR) and Lasso

1) Support Vector Regression (SVR): The Support Vector Machine (SVM) is a widely used supervised learning method. In this paper, we focus on the SVMs for regression, in order to predict the performance ratings from interview features. Suppose we are given a training

TABLE I

LIST OF ASSESSMENT QUESTIONS ASKED TO AMAZON MECHANICAL TURK WORKERS.

Traits	Description
Overall Rating	The overall performance rating.
Recommend Hiring	How likely is he to get hired?
Engagement	Did he use engaging voice?
Excitement	Was he excited?
Eye Contact	Did he maintain proper eye contact?
Smile	Did he smiled appropriately?
Friendliness	Did he seem friendly?
Speaking Rate	Did he maintain a good speaking rate?
No Fillers	Did he use too many filler words? (1 = too many, 7 = no filler words)
Paused	Did he pause appropriately?
Authentic	Did he seem authentic?
Calm	Did he appear to be calm?
Structured Answer	Were his answers structured?
Focused	Did he seem focused?
Not Stressed	Was he stressed? (1 = too stressed, 7 = not stressed)
Not Awkward	Did he seem awkward? (1 = too awkward, 7 = not awkward)

TABLE II

LIST OF PROSODIC FEATURES AND THEIR BRIEF DESCRIPTIONS

Prosodic Feature	Description
Energy	Mean spectral energy.
F0 MEAN	Mean F0 frequency.
F0 MIN	Minimum F0 frequency.
F0 MAX	Maximum F0 frequency.
F0 Range	Difference between F0 MAX and F0 MIN.
F0 SD	Standard deviation of F0.
Intensity MEAN	Mean vocal intensity.
Intensity MIN	Minimum vocal intensity .
Intensity MAX	Maximum vocal intensity .
Intensity Range	Difference between max and min intensity.
Intensity SD	Standard deviation.
F1, F2, F3 MEAN	Mean frequencies of the first 3 formants: F1, F2, and F3.
F1, F2, F3 SD	Standard deviation of F1, F2, F3.
F1, F2, F3 BW	Average bandwidth of F1, F2, F3.
F2/F1 MEAN	Mean ratio of F2 and F1.
F3/F1 MEAN	Mean ratio of F3 and F1.
F2/F1 SD	Standard deviation of F2/F1.
F3/F1 SD	Standard deviation of F3/F1.
Jitter	Irregularities in F0 frequency.
Shimmer	Irregularities in intensity.
Duration	Total interview duration.
% Unvoiced	Percentage of unvoiced region.
% Breaks	Average percentage of breaks.
maxDurPause	Duration of the longest pause.
avgDurPause	Average pause duration.

TABLE III

LIWC LEXICAL FEATURES USED IN OUR SYSTEM.

LIWC Category	Examples
I	*I, I'm, I've, I'll, I'd,* etc.
We	*we, we'll, we're, us, our,* etc.
They	*they, they're, they'll, them,* etc.
Non-fluencies	words introducing non-fluency in speech, e.g., *uh, umm, well.*
PosEmotion	words expressing positive emotions, e.g., *hope, improve, kind, love.*
NegEmotion	words expressing negative emotions, e.g., *bad, fool, hate, lose.*
Anxiety	*nervous, obsessed, panic, shy,* etc.
Anger	*agitate, bother, confront, disgust,* etc.
Sadness	*fail, grief, hurt, inferior,* etc.
Cognitive	*cause, know, learn, make, notice,* etc.
Inhibition	*refrain, prohibit, prevent, stop,* etc.
Perceptual	*observe, experience, view, watch,* etc.
Relativity	*first, huge, new,* etc.
Work	*project, study, thesis, university,* etc.
Swear	Informal and swear words.
Articles	*a, an, the,* etc.
Verbs	common English verbs.
Adverbs	common English adverbs.
Prepositions	common prepositions.
Conjunctions	common conjunctions.
Negations	*no, never, none, cannot, don't,* etc.
Quantifiers	*all, best, bunch, few, ton, unique,* etc.
Numbers	words related to number, e.g., *first, second, hundred,* etc.

data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N))\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is a d-dimensional feature vector for the i^{th} interview in the training set. For each feature vector $\mathbf{x}_i$, we have an associated value $y_i \in \mathbb{R}_+$ denoting the interview rating. Our goal is to learn the optimal weight vector $\mathbf{w} \in \mathbb{R}^d$ and a scalar bias term $b \in \mathbb{R}$ such that the predicted value for the feature vector $\mathbf{x}$ is: $\hat{y} = \mathbf{w}^T\mathbf{x} + b$. We minimize the following objective function:

$$\begin{aligned} \underset{\mathbf{w},\xi_i,\hat{\xi}_i,b}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + \hat{\xi}_i) \\ \text{subject to} \quad & y_i - \mathbf{w}^T\mathbf{x}_i - b \leq \epsilon + \xi_i, \ \forall i \\ & \mathbf{w}^T\mathbf{x}_i + b - y_i \leq \epsilon + \hat{\xi}_i, \ \forall i \\ & \xi_i, \hat{\xi}_i \geq 0, \ \forall i \end{aligned} \tag{5}$$

The $\epsilon \geq 0$ is the precision parameter specifying the amount of deviation from the true value that is allowed, and $(\xi_i, \hat{\xi}_i)$ are the slack variables to allow deviations larger than ϵ. The tunable parameter $C > 0$ controls the tradeoff between goodness of fit and generalization to new data. The convex optimization problem is often solved by maximizing the corresponding dual problem. In order to analyze the relative weights of different features, we transform it back to the primal problem and obtain the optimal weight vector $\mathbf{w}^*$ and bias term b^*. The relative importance of the j^{th} feature can be interpreted by the associated weight magnitude $|w_j^*|$.

2) Lasso: The Lasso regression method aims to minimize the residual prediction error in the presence of an L_1 regularization function. Using the same notation as the previous section, let the training data be $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N))\}$. Let our linear predictor be of the form: $\hat{y} = \mathbf{w}^T\mathbf{x} + b$. The Lasso method estimates the optimal $\mathbf{w}$ and b by minimizing the following objective function:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T\mathbf{x}_i - b\right)^2 \\ \text{subject to} \quad & \|\mathbf{w}\|_1 \leq \lambda \end{aligned} \tag{6}$$

where $\lambda > 0$ is the regularization constant, and $\|\mathbf{w}\|_1 = \sum_{j=1}^{d} |w_j|$ is the L_1 norm of $\mathbf{w}$. The L_1 regularization is known to push the coefficients of the irrelevant features down to zero, thus reducing the predictor variance. We control the amount of sparsity in the weight vector $\mathbf{w}$ by tuning the regularization constant λ.

APPENDIX
LIST OF MOST IMPORTANT FEATURES

For both SVR and Lasso models, we sort the features by the magnitude of their weights and examine the top twenty features (excluding the topic features). These features and their weights are listed in Table IV and Table V for SVR and Lasso respectively.

REFERENCES

[1] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *The Journal of Machine Learning Research*, vol. 99, pp. 1297–1322, 2010.

TABLE IV

FEATURE ANALYSIS USING THE SVR MODEL. WE ARE LISTING THE TOP TWENTY FEATURES ORDERED BY THEIR WEIGHT MAGNITUDE. WE HAVE EXCLUDED THE TOPIC FEATURES FOR THE EASE OF INTERPRETATION.

Overall		Recommend Hiring		Excited		Engagement		Friendly	
avgBand1	-0.116	wpsec	0.136	avgBand1	-0.153	avgBand1	-0.166	smile	0.258
wpsec	0.104	avgBand1	-0.132	diffIntMaxMin	0.129	intensityMax	0.162	mean pitch	0.169
Quantifiers	0.087	Fillers	-0.129	f3STD	-0.125	intensityMean	0.142	f3STD	-0.116
avgDurPause	-0.087	percentUnvoiced	-0.116	smile	0.123	diffIntMaxMin	0.14	intensityMax	0.101
Fillers	-0.086	smile	0.105	mean pitch	0.121	wpsec	0.13	f1STD	-0.095
upsec	0.083	upsec	0.099	wpsec	0.121	avgBand2	-0.122	diffIntMaxMin	0.094
percentUnvoiced	-0.082	PercentBreaks	-0.097	intensityMax	0.119	f1STD	-0.113	intensityMean	0.093
smile	0.082	avgDurPause	-0.095	f1STD	-0.113	f2STDf1	0.104	Adverbs	0.09
Relativity	0.078	f3meanf1	0.082	percentUnvoiced	-0.111	f3meanf1	0.102	shimmer	-0.087
f3meanf1	0.076	f1STD	-0.082	intensityMean	0.109	f3STD	-0.099	wpsec	0.085
maxDurPause	-0.073	intensityMean	0.081	nod	0.107	Quantifiers	0.094	percentUnvoiced	-0.083
PercentBreaks	-0.071	nod	0.079	PercentBreaks	-0.106	upsec	0.092	PercentBreaks	-0.082
f1STD	-0.071	Quantifiers	0.078	intensitySD	0.099	intensitySD	0.089	fmean3	0.079
Positive emotion	-0.066	maxDurPause	-0.074	f2STDf1	0.091	percentUnvoiced	-0.088	max pitch	0.077
f2STDf1	0.064	Prepositions	0.072	f3meanf1	0.09	smile	0.086	I	-0.075
Prepositions	0.061	Positive emotion	-0.072	Adverbs	0.09	PercentBreaks	-0.085	avgBand1	-0.072
intensityMean	0.059	Articles	0.071	Non-fluencies	-0.083	shimmer	-0.081	upsec	0.072
uc	0.059	f2meanf1	0.069	f2meanf1	0.082	f2meanf1	0.075	nod	0.065
f3STD	-0.057	f3STD	-0.068	avgBand2	-0.082	Adverbs	0.074	diffPitchMaxMin	0.064
wc	0.057	uc	0.067	wc	0.079	max pitch	0.073	We	0.06

TABLE V

FEATURE ANALYSIS USING THE LASSO MODEL. WE ARE LISTING THE TOP TWENTY FEATURES ORDERED BY THEIR WEIGHT MAGNITUDE. WE HAVE EXCLUDED THE TOPIC FEATURES FOR THE EASE OF INTERPRETATION.

Overall		Recommend Hiring		Excited		Engagement		Friendly	
avgBand1	-0.562	avgBand1	-0.585	avgBand1	-0.722	intensityMax	0.697	smile	0.516
wpsec	0.313	wpsec	0.417	intensityMax	0.27	avgBand1	-0.692	intensityMax	0.444
Fillers	-0.219	Fillers	-0.366	wpsec	0.262	wpsec	0.36	mean pitch	0.324
percentUnvoiced	-0.089	percentUnvoiced	-0.158	mean pitch	0.161	mean pitch	0.128	wpsec	0.166
Quantifiers	0.059	smile	0.111	smile	0.157	shimmer	-0.081	f3STD	-0.137
smile	0.056	Quantifiers	0.051	diffIntMaxMin	0.152	smile	0.077	diffIntMaxMin	0.057
Relativity	0.019	Articles	0.018	wc	0.098	intensityMean	0.066	avgBand1	-0.039
PercentBreaks	-0.005	max pitch	0.014	f3STD	-0.089	upsec	0.044	f1STD	-0.033
avgDurPause	-0.003	nod	0.01	percentUnvoiced	-0.081	Quantifiers	0.037	Cognitive	0.021
Conjunctions	0.003	wc	0.007	nod	0.057	PercentBreaks	-0.026	Adverbs	0.017
f3meanf1	0.002	mean pitch	0.006	PercentBreaks	-0.02	percentUnvoiced	-0.023	intensityMean	0.016
maxDurPause	-0.002	Conjunctions	0.005	shimmer	-0.009	f3STD	-0.021	Sadness	0.01
Positive emotion	-0.001	fpsec	-0.005	Cognitive	0.006	Conjunctions	0.005	f2STDf1	0.008
mean pitch	0.001	avgDurPause	-0.004	intensityMean	0.004	diffIntMaxMin	0.004	max pitch	0.005
Prepositions	0.001	Perceptual	-0.004	Quantifiers	0.004	max pitch	0.003	shimmer	-0.004
f1STD	-0.001	f3meanf1	0.003	Adverbs	0.002	f1STD	-0.003	fpsec	0.002
fpsec	-0.0	Relativity	0.002	Non-fluencies	-0.002	avgBand2	-0.002	percentUnvoiced	-0.0
upsec	0.0	PercentBreaks	-0.001	f3meanf1	0.001	Cognitive	0.002	I	-0.0
f3STD	-0.0	intensityMean	0.001	max pitch	0.001	fmean3	0.001	We	0.0
f2STDf1	0.0	Prepositions	0.001	avgBand2	-0.001	f3meanf1	0.001	Positive emotion	0.0