Automated Analysis and Prediction of Job Interview Performance

Iftekhar Naim, *Student Member, IEEE*, Md. Iftekhar Tanveer, *Student Member, IEEE*, Daniel Gildea, and Mohammed Ehsan Hoque, *Member, IEEE*

Abstract—We present a computational framework for automatically quantifying verbal and nonverbal behaviors in the context of job interviews. The proposed framework is trained by analyzing the videos of 138 interview sessions with 69 internship-seeking undergraduates at the Massachusetts Institute of Technology (MIT). Our automated analysis includes facial expressions (e.g., smiles, head gestures, facial tracking points), language (e.g., word counts, topic modeling), and prosodic information (e.g., pitch, intonation, and pauses) of the interviewees. The ground truth labels are derived by taking a weighted average over the ratings of nine independent judges. Our framework can automatically predict the ratings for interview traits such as excitement, friendliness, and engagement with correlation coefficients of 0.70 or higher, and can quantify the relative importance of prosody, language, and facial expressions. By analyzing the relative feature weights learned by the regression models, our framework recommends to speak more fluently, use fewer filler words, speak as "we" (versus "I"), use more unique words, and smile more. We also find that the students who were rated highly while answering the first interview question were also rated highly overall (i.e., first impression matters). Finally, our MIT Interview dataset is available to other researchers to further validate and expand our findings.

Index Terms—Nonverbal behavior prediction, job interviews, multimodal interactions, regression

1 Introduction

Analysis of non-verbal behavior to predict the outcome of a social interaction has been studied for many years in different domains, with predictions ranging from marriage stability based on interactions between newlywed couples [1], [2], to patient satisfaction based on doctor-patient interaction [3], to teacher evaluation by analyzing classroom interactions between a teacher and the students [4]. However, many of these prediction frameworks were based on manually labeled behavioral patterns by trained coders, according to carefully designed coding schemes. Manual labeling of nonverbal behaviors is laborious and time consuming, and therefore often does not scale with large amounts of data. As a scalable alternative, several automated prediction frameworks have been proposed based on low-level behavioral features, automatically extracted from larger datasets. Due to the challenges of collecting and analyzing multimodal data, most of these automated methods focused on a single modality of interaction [5], [6], [7], [8]. In this paper, we address the challenge of automated understanding of multimodal human interactions, including facial expression, prosody, and language. We focus on predicting attributes of social interactions in the context of job interviews for college students, which is an exciting and relatively less explored domain.

Job interviews are ubiquitous and play inevitable and important roles in our life and career. Over many years, social psychologists and career coaches have accumulated knowledge and guidelines for success in job interviews [9], [10], [11]. Studies in social psychology have shown that smiling, using a louder voice, and maintaining eye contact contribute positively to our interpersonal communications [9], [11]. These guidelines are largely based on intuition, experience, and studies involving manual encoding of nonverbal behaviors on a limited amount of data [9]. Automated data-driven quantification of both verbal and non-verbal behaviors simultaneously has not been explored in the context of job interviews until very recently. In our prior work [12], we quantified the determinants of a successful job interview using a computational prediction framework based on automatically extracted features, which takes both verbal speech and non-verbal behaviors into account. In this article, we present an improved system by including additional facial features, and provide more comprehensive experiments, results, and analysis.

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]".*

- *I. Naim and D. Gildea are with the Department of Computer Science, University of Rochester, Rochester, NY 14627. E-mail: {inaim, gildea}@cs.rochester.edu.*
- *Md. I. Tanveer is with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627. E-mail: itanveer@cs.rochester.edu.*
- *Md. E. Hoque is with the Department of Computer Science, and the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627. E-mail: mehoque@cs.rochester.edu.*

Manuscript received 13 Apr. 2015; revised 3 June 2016; accepted 4 Aug. 2016. Date of publication 27 Sept. 2016; date of current version 6 June 2018. Recommended for acceptance by L.-P Morency. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TAFFC.2016.2614299

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 did almost 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. [looked away from the interviewer]".

Who do you think received higher ratings?

Most would agree that the first interviewee, John, provided the 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 cannot quantify how the combination of these behaviors affects our interpersonal communications. Previous research [13] shows that the style of speaking, prosody, facial expression, and language reflect valuable information about one's personality and mental states. 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 the audio-visual recordings of 138 interview 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 infer the relative influence of individual interview questions and/or different temporal phases (e.g., beginning, ending) of job interviews?
- 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 ratings of job interviews, given the audio-visual recordings. The proposed prediction framework (Fig. 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 job interview process. Our system is capable of predicting the overall rating of a job interview with a correlation coefficient $r > 0.62$ and AUC = 0.77 (random chance baseline is 0.50) on average. We can also predict different social traits such as engagement, excitement, and friendliness with even higher accuracy ($r \geq 0.70$, AUC > 0.80). Furthermore, we investigate the relative weights of the individual verbal and non-verbal features learned by our regression models, and quantify their relative importance in the context of job interviews. With the exceptions of transcribing interviewers' speech to text and collecting ground truth ratings (which were performed by Amazon Mechanical Turk workers), all the other modules in our pipeline are fully automated. With the remarkable recent advances in speech recognition [14], we expect our results to carry over to automatically generated speech-to-text transcriptions. Our prediction model can be integrated with the existing automated interview coaching systems, such as MACH [15], to provide more intelligent and quantitative feedback. 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 a specific job.

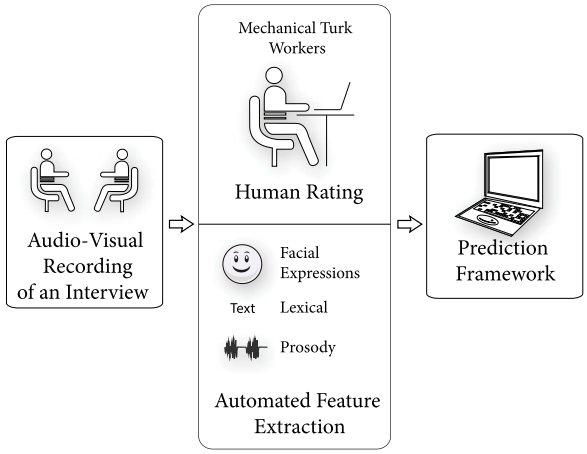


Fig. 1. Framework of Analysis. Mechanical Turk workers rated interviewee performance by watching videos of job interviews. Various features were extracted from those videos. A framework was built to predict Turker's rating and to gain insight into the characteristics of a good interview.

Parts of the research included in this article have been presented in [12]. In this paper, we extend our previous work by including novel research questions and incorporating more comprehensive features, experiments, and analysis. We study the relative influence of individual interview questions and different temporal phases of interviews on the overall job interview ratings and show that the questions asked during the beginning and end of the interviews tend to have a higher impact. Furthermore, we perform new experiments to study the relative influence of facial expression, prosody, and language, and their different combinations. We compare our prediction models with the ratings provided by one of the nine judges chosen at random. Our models significantly outperform this strong baseline for several different traits (especially for overall performance, hiring recommendation, excitement, engagement, and friendliness).

The remaining structure of the article follows. In Section 2, we discuss the background research on automated quantification of multimodal nonverbal behaviors. Section 3 describes the interview dataset and the data annotation process via Mechanical Turk. A detailed discussion of the proposed computational framework, feature extraction, and automated prediction is presented in Section 4. We present our detailed results in Section 5. Finally, we conclude with our findings and discuss our future work in Section 6.

2 Background Research

In this section, we discuss existing relevant work on nonverbal behavior prediction using automatically extracted features. We particularly focus on the social cues that have been shown to be relevant to job interviews and face-to-face

interactions [9]. We also discuss previous research on automated conversational systems for job interviews, which is one of the potential applications we envision for the proposed prediction framework.

2.1 Nonverbal Behavior Recognition

Nonverbal behaviors are subtle, fleeting, subjective, and sometimes even contradictory. Even a simple facial expression such as a smile can have different meanings, e.g., delight, rapport, sarcasm, and even frustration [16]. Edward Sapir, in 1927, referred to non-verbal behavior as "an elaborate and secret code that is written nowhere, known by none, but understood by all" [17]. Despite years of research, nonverbal behavior prediction remains a challenging problem. Gottman et al. [1], [2] studied verbal and non-verbal interactions between newlywed couples and developed mathematical models to predict marriage stability and chances of divorce. For example, they found that the greatest predictor of divorce is contempt, which must be avoided for a successful marriage. Hall et al. [3] studied the nonverbal cues in doctor-patient interaction and showed that doctors who are more sensitive to nonverbal skills received higher ratings of service during patient visits. Ambady et al. [4] studied the interactions of teachers with students in a classroom and proposed a framework for predicting teachers' evaluations based on short clips of interactions. However, these prediction frameworks were based on manually labeled behavioral patterns. Manually labeling nonverbal behaviors is laborious and time consuming, and is often not scalable to large amounts of data.

To allow for the analysis of larger datasets of social interactions, several automated prediction frameworks have been proposed. Due to the challenges of collecting and analyzing multimodal data, most of the existing automated prediction frameworks focus on a single behavioral modality, such as prosody [8], [18], [19], facial expression [6], gesture [7], [20], and word usage pattern [21]. Analysis based on a single modality is likely to overlook many critical non-verbal behaviors, and hence there has been a growing interest in analyzing social behaviors in more than a single modality.

Ranganath et al. [22], [23] studied social interactions in speed-dates using a combination of prosodic and linguistic features. The analysis is based on the SpeedDate corpus, a spoken corpus of approximately 1000 4-min-speed-dates, where each participant rated his/her date in terms of four different conversational styles (awkwardness, assertiveness, flirtatiousness, and friendliness) on a ten point Likert scale. Given the speech data, Ranganath et al. proposed a computational framework for predicting these four conversational styles using prosodic and linguistic features only, while ignoring facial expressions. Stark et al. [24] were able to reliably predict the nature of a telephone conversation (business versus personal, familiar versus unfamiliar) using the lexical and prosodic features extracted from as few as 30 words of speech at the beginning of the conversation. Kapoor et al. [13] and Pianesi et al. [25] proposed systems to recognize different social and personality traits by exploiting only prosody and visual features. Sanchez et al. [26] proposed a system for predicting eleven different social moods (e.g., surprise, anger, happiness) from YouTube video monologues. Related studies on YouTube videos have been explored in [27], [28], [29]. Nguyen and Gatica-Perez [30] studied YouTube video resumes, and explored how nonverbal behaviors shown in these videos influence hiring decisions. However, the social dynamics of YouTube monologues are different from face to face interactions.

Perhaps the most relevant, Nguyen et al. [31] proposed a computational framework to predict the hiring decision using nonverbal behavioral cues extracted from a dataset of 62 interview videos. Nguyen et al. considered only nonverbal cues, and did not include verbal content in their analysis. Our work extends the current state-of-the-art and generates new knowledge by incorporating three different modalities (prosody, language, and facial expressions), and fifteen 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 success in job interviews (Section 5.2.3).

2.2 Social Coaching for Job Interviews

Several automated systems have been proposed for coaching the necessary social skills to succeed in job interviews [15], [32], [33]. Hoque et al. [15] developed MACH (My Automated Conversation coacH), which allows users to improve social skills by interacting with a virtual agent. The MACH system records videos of the user using a webcam and a microphone, and provides feedback regarding several low level behavioral patterns, e.g., average smile intensity, pause duration, speaking rate, pitch variation, etc.

Anderson et al. [32] proposed an interview coaching system, TARDIS, which presents the training interactions as a scenario-based "serious game". The TARDIS framework incorporates a sub-module named NovA (NonVerbal behavior Analyzer) [33] that can recognize several lower level social cues: *hands-to-face*, *looking away*, *postures*, *leaning forward/backward*, *gesticulation*, *voice activity*, *smiles*, and *laughter*. Using these manually annotated ground truth social cues as features, NovA trains a Bayesian Network that can infer higher-level mental traits (e.g., stressed, focused, engaged, etc.). However, prediction of higher-level traits using automatically extracted features remains part of their future work.

Our framework (1) quantifies the relative influences of different low-level features on the interview outcome, (2) learns regression models to predict interview ratings and the likelihood of hiring using automatically extracted features, and (3) predicts several other high-level personality traits such as engagement, friendliness, and excitement. One of our objectives is to extend the existing automated conversation systems by providing feedback on the overall interview performance and additional high-level personality traits.

3 Dataset Description

We used the *MIT Interview Dataset* [15], which consists of 138 audio-visual recordings of mock interviews with internship-seeking students from Massachusetts Institute of Technology (MIT). The total duration of our interview videos is nearly 10.5 hours (on average, 4.7 minutes per interview, for

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.

138 interview videos). To our knowledge, this is the largest collection of interview videos conducted by professional counselors under realistic settings. Our dataset and the relevant features are made publicly available[1] for research puproses. The following sections provide a brief description of the data collection and ground truth labeling.

3.1 Data Collection

3.1.1 Study Setup

The mock interviews were conducted in a room equipped with a desk, two chairs, and two wall-mounted cameras, as shown in Fig. 2. The two cameras with microphones were used to capture the facial expressions and the audio conversations during the interview.

3.1.2 Participants

Initially, 90 MIT juniors participated in the mock interviews. All participants were native English speakers. The interviews were conducted by two professional MIT career counselors who had over five years of experience. For each participant, two rounds of mock interviews were conducted: before and after interview intervention. For the details of interview intervention, please see [15]. Each individual received $50 for participating. Furthermore, as an incentive for the participants, we promised to forward the resume of the top 5 percent candidates to several sponsor organizations (Deloitte, IDEO, and Intuit) for consideration for summer internships. We chose sponsor organizations which are not directly tied to any specific major. After the data collection, 69 (26 male, 43 female) of the 90 initial participants permitted the use of their video recordings for research purposes.

3.1.3 Procedure

During each interview session, the counselor asked interviewees five different questions, which were recommended by the MIT Career Services. These five questions were presented in the following order by the counselors to the participants:

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?

No job description was given to the interviewees. The five questions were chosen to assess the interviewee's behavioral and social skills. The interviewers rated the performances of the interviewees by answering 16 assessment questions on a seven point Likert scale. We list these questions in Table 1. These questions to the interviewers were selected to evaluate the overall performance and behavioral traits of the interviewees. The first two questions – "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").

TABLE 1
List of Questions Asked to Mechanical Turk Workers

Traits	Description
Overall Rating	The overall performance rating.
Recommend Hiring	How likely is he to be hired?
Engagement	Did he use engaging tone?
Excitement	Did he seem excited?
Eye Contact	Did he maintain proper eye contact?
Smile	Did he smile 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 calm?
Focused	Did he seem focused?
Structured Answers	Were his answers structured?
Not Stressed	Was he stressed? (1 = too stressed, 7 = not stressed)
Not Awkward	Did he seem awkward? (1 = too awkward, 7 = not awkward)

3.2 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 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. The Mechanical Turkers used the same questionnaire to assess the ratings as listed in Table 1. Apart from being less affected by bias, the Mechanical Turk workers could pause and replay the video, allowing them to rate more thoroughly. The Turkers' ratings are more likely to be similar to "audience" ratings, as opposed to "expert ratings".

1. www.cs.rochester.edu/hci/currentprojects.php?proj=interview

In order to collect ground truth ratings for interviewee performances, 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. Only Turkers located in United States and having a HIT Approval Rate higher than 95 percent were allowed to participate. We automatically estimated the quality of individual workers using an EM-style optimization algorithm, and estimated a weighted average of their scores as the ground truth ratings, which were used in our prediction framework.

One of our objectives was to model the temporal relationships among the individual interview questions and the overall ratings. To accomplish this, we performed a second phase of labeling. We hired a different set of five Turkers for rating the performances of an interviewee in each of the five interview questions separately. This was done by splitting each interview video into five different segments, where each segment corresponds to one of the interview questions. The video segments were shuffled so that each Turker would rate the segments in a random order. These per-question ratings were used only to analyze the temporal variation in the ratings and measure how the temporal order of the questions correlates with the ratings for entire interview.

4 Prediction Framework

For the prediction framework, we automatically extracted various features from the videos of the interviews. Then we trained two regression algorithms - SVR and LASSO. The objective of this training is twofold: first, to predict the Turker's ratings on the overall performance and each behavioral trait, and second, to quantify and gain meaningful insights on the relative importance of each modality and the interplay among them.

4.1 Feature Extraction

We collected three types of features for each interview video: (1) prosodic features, (2) lexical features, and (3) facial features. We selected these features to reflect the behaviors that have been shown to be relevant in job interviews (e.g., smile, intonation, language content, etc.) [9], and also based on the past literature on automated social behavior recognition [19], [22], [23], [26]. 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 (e.g., "uh", "umm", "like") in the transcriptions. Our lexical features were extracted from these transcripts. We also collected a wide range of prosodic and facial features.

4.1.1 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 [8], [18], [19]. To distinguish between the speech of the interviewer and the interviewee, we manually annotated the beginning and end of each of the interviewee's answers. We extracted and analyzed prosodic features of the interviewee's speech. Each prosodic feature is first collected over an interval corresponding to a single answer by the interviewee, and then averaged over all her/his five answers. We used the open-source speech analysis tool PRAAT [34] for prosody analysis.

TABLE 2
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.

The important prosodic features include pitch information, vocal intensities, characteristics of the first three formants, and spectral energy, which have been reported to reflect our social traits [18]. To reflect the vocal pitch, we extracted the mean and standard deviation 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 three formants. Additionally, we collected several other prosodic features such as pause duration, percentage of unvoiced frames, jitter (irregularities in pitch), shimmer (irregularities in vocal intensity), percentage of breaks in speech, etc. Table 2 shows the complete list of prosodic features.

4.1.2 Lexical Features

Lexical features can provide valuable information regarding the interview content and the interviewee's personality. One of the most commonly used lexical features is the unigram counts for each individual word. However, treating unigram counts as features 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) [35]. The

TABLE 3
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.

LIWC categories include words describing negative emotions (sad, angry, etc.), positive emotions (happy, kind, etc.), different function word categories (articles, quantifiers, pronouns, etc.), and various content categories (e.g., anxiety, insight). We apply a *greedy backward elimination feature selection* [36], which starts with all the LIWC features in the LIWC-2007 inventory, and sequentially removes one feature at a time such that its removal results in maximum accuracy gain in cross-validation. We continue this process iteratively until any further feature removal result in a noticeable decrease in accuracy, and thus select 23 LIWC features. The LIWC categories correlate with various psychological traits, and often provide indications about our personality and social skills [21]. Many of these categories are intuitively related to interview performance. Table 3 shows the complete list of the LIWC features used in our experiments.

Although the hand coded LIWC lexicon has proven to be useful for modeling many different social behaviors, the lexicon is predefined and may not cover many important aspects of job interviews. To address this challenge, we aimed to automatically learn a lexicon from the interview dataset. We apply the Latent Dirichlet Allocation (LDA) [37] method 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. Similar ideas have been exploited by Ranganath et al. [22], [23] for modeling social traits in the speed dating dataset, but they used deep auto-encoders [38] instead of LDA.

Finally, we collected additional lexical features related to our linguistic and speaking skills (listed in Table 4). 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.

TABLE 4
Additional Features Related to Speaking Rate and Fluency

feature Name	Description
wpsec	Words per second.
upsec	Unique words per second.
fpsec	Filler words per second.
wc	Total number of words.
uc	Total number of unique words.

4.1.3 Facial Features

We extracted facial features for the interviewees from each frame in the video. First, faces were detected using the Shore [39] framework. We trained a classifier to distinguish between neutral and smiling faces. The classifier is trained using the AdaBoost algorithm. The classifier output is normalized in the range [0,100], where 0 represents no smile, and 100 represents full smile. Finally, we averaged the smile intensities from individual frames, and used this as a feature in our model. We also extracted head gestures such as nods and shakes as explained in [15].

In addition to the smile intensity and head gestures (nod and shake), we also extracted a number of other facial features using a Constrained Local Model (CLM) [40] based face tracker,[2] as illustrated in Fig 3. The face tracker detects 66 interest points on a face image. It works by fitting the following parametric shape model [40], [41]

$$\mathbf{x}_i = s\mathbf{R}(\mathbf{x}_i + \mathbf{\Psi_i q}) + \mathbf{t}, \tag{1}$$

where $\mathbf{x}_i$ is the coordinate of ith interest point and $\mathbf{x}_i$ denotes its mean location pre-trained from a large collection of hand-labeled training images. $\mathbf{\Psi_i}$ denotes the bases of local variations for the ith interest point. Each element of the vector $\mathbf{q}$ represents a coefficient corresponding to a basis of local variation. The parameters $s, \mathbf{R}$, and t corresponds to the global transformations associated with scaling, rotation, and translation respectively. The face tracker adjusts the model parameters $p = \{s, \mathbf{R}, \mathbf{q}, \mathbf{t}\}$ so that each of the mean interest points ($\mathbf{x}_i$) fits best to its corresponding point ($\mathbf{x}_i$) on the test face.

While extracting features from these tracked interest points, we want to disregard the global transformations (translation, rotation, and scaling), and consider only the local transformations, which provide useful information regarding our facial expressions. After the face tracker converges to an optimal estimate of the parameters, we recalculate each of the interest points $\mathbf{x}_i$ by applying the local transformations only, while disregarding the global transformations ($s, \mathbf{R}$, and t). Mathematically, we calculate the following shape model from the optimal parameters obtained from the face tracker

$$\hat{\mathbf{x}}_i = (\mathbf{x}_i + \mathbf{\Psi_i q}), \tag{2}$$

2. https://github.com/kylemcdonald/FaceTracker

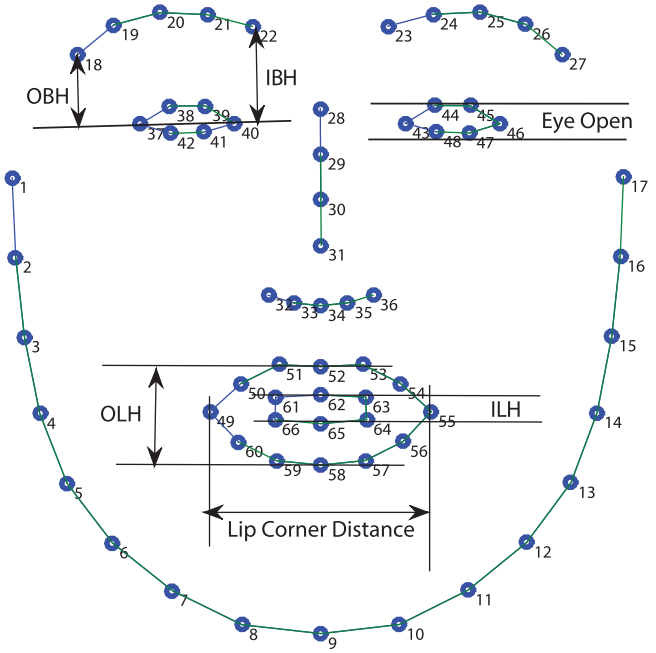


Fig. 3. Illustration of facial features: OBH (outer eye-brow height), IBH (inner eye-brow height), OLH (outer lip height), ILH (inner lip height), eye opening, and LipCDT (lip corner distance).

Once we find $\hat{\mathbf{x}}_i$, we calculate the distances between the corresponding interest points to find out the features OBH (outer eye-brow height), IBH (inner eye-brow height), OLH (outer lip height), and ILH (inner lip height), eye opening, and LipCDT (lip corner distance), as illustrated in Fig. 3. By disregarding the global transformation parameters, the extracted facial features are invariant to global translations, rotations, and scaling variations. In addition to the features shown in Fig. 3, we separately incorporated three head pose features (Pitch, Yaw and Roll), based on the corresponding elements of the rotation matrix $\mathbf{R}$.

4.1.4 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, which allows treating all the features uniformly.

4.2 Ground Truth Ratings and Turker Quality Estimation

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 models [42] that treat each Turker's reliability and ground truth ratings as latent parameters, estimate their values using an EM-style iterative optimization technique.

Let us assume an input training dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$ containing N feature vectors $\mathbf{x}_i$ (one for each interview video), for which the ground truth label y_i is not known. Instead 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 the reliability of each worker (λ_j).

To simplify the estimation problem, we assume the Turkers' ratings to be 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:

$$Pr[y_i^j | y_i, \lambda_j] = \mathcal{N}(y_i^j | y_i, 1/\lambda_j), \tag{3}$$

where λ_j is the unknown inverse-variance and the measure of reliability for the jth Turker. By taking logarithm on both side and ignoring constant terms, we get the log-likelihood function:

$$L = \sum_{i=1}^{N} \sum_{j=1}^{K} \left[\frac{1}{2} \log \lambda_j - \frac{\lambda_j}{2} (y_i^j - y_i)^2 \right], \tag{4}$$

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:

$$y_i = \frac{\sum_{j=1}^{K} \lambda_j y_i^j}{\sum_{j=1}^{K} \lambda_j}, \tag{5}$$

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

$$\lambda_j = \frac{\sum_{i=1}^{N} (y_i^j - y_i)^2}{N}, \tag{6}$$

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.

4.3 Score Prediction from Extracted Features

Using the features described in the previous section, we train regression models to predict the interview scores. We also train models to predict other interview-specific traits such as excitement, friendliness, engagement, awkwardness, etc. We experimented with many different regression models: Support Vector Machine Regression (SVR) [43], Lasso [44], L_1 Regularized Logistic Regression, Gaussian Process Regression, etc. We will only discuss SVR and Lasso, which achieved the best results with our dataset.

4.3.1 Support Vector Regression

The Support Vector Machine (SVM) is a widely used supervised learning method for classification. 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 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 ith 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:

$$\begin{aligned} \underset{\mathbf{w},\xi_i,\hat{\xi}_i,b}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + \hat{\xi}_i) \\ \text{subject to} \quad & y_i - \mathbf{w}^T\mathbf{x}_i - b \leq \epsilon + \xi_i, \ \forall i \\ & \mathbf{w}^T\mathbf{x}_i + b - y_i \leq \epsilon + \hat{\xi}_i, \ \forall i \\ & \xi_i, \hat{\xi}_i \geq 0, \ \forall i \end{aligned} \tag{7}$$

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 jth feature can be interpreted by the associated weight magnitude $|w_j^*|$.

4.3.2 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:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & \sum_{i=1}^{N}\left(y_i - \mathbf{w}^T\mathbf{x}_i - b\right)^2 \\ \text{subject to} \quad & \|\mathbf{w}\|_1 \leq \lambda, \end{aligned} \tag{8}$$

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.

5 Results

We organize our results in two sections. First, we analyze the ratings by Mechanical Turk workers (Section 5.1). The quality and reliability of Turkers' ratings are assessed by observing how well the Turkers agree with each other (Section 5.1.1). In addition, we identify which traits are important to succeed in job interviews by measuring the correlations of the ratings for individual traits with the overall ratings (Section 5.1.2). Furthermore, we examine the correlations between the ratings for individual video segments with that for the entire videos. This allowed us to evaluate the temporal patterns in job interviews (Section 5.1.3).

In Section 5.2, we present the prediction accuracies for the trained regression models (SVR and Lasso) based on automatically extracted features, and analyze the relative influence of different modalities and features on prediction accuracy.

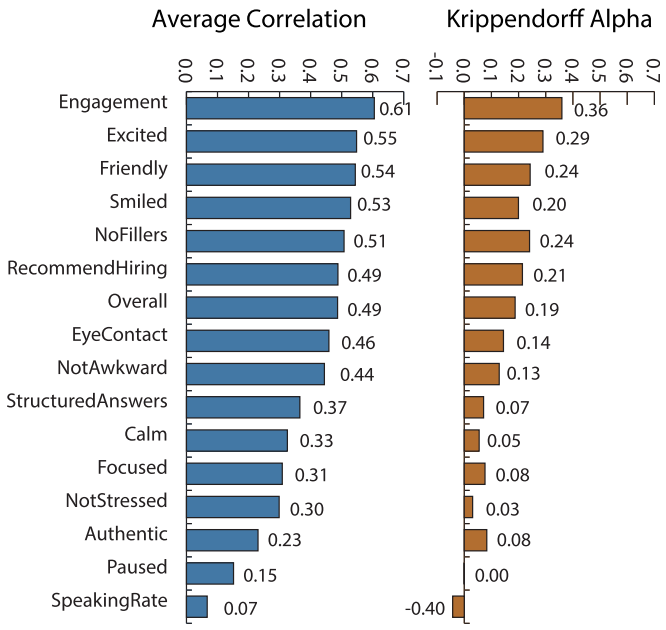


Fig. 4. The inter-rater agreement among the turkers, measured by the Krippendorff's Alpha (varies in the range $[-1, 1]$) and the average one-vs-rest correlation of their ratings (range $[-1, 1]$).

5.1 Analysis of Mechanical Turk Dataset

5.1.1 Inter-Rater Agreement

To assess the quality of the ratings, we calculate Krippendorff's Alpha [45] for each trait. In this case, Krippendorff's Alpha is more meaningful than the frequently used Fleiss' Kappa [46], 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 $[-1, 1]$, with 1 being the perfect agreement and -1 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. Fig. 4 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. This may be because the Turkers were not in a position to judge those categories with the video data only. The high variability among the Turkers' ratings illustrates the subjective nature of these ratings, and justifies our decision of collecting multiple ratings.

5.1.2 Correlation among the Behavioral Traits

We are interested in identifying the traits that correlate highly with overall ratings. This knowledge can help interviewees understand the most important behavioral traits in job interviews. We plot the mutual information and correlation between various ratings given by the Mechanical Turk workers and the overall rating of the interviewee performance in Fig. 5.

The first bar in Fig. 5 represents whether the rater will recommend hiring the interviewee. It is another form of the overall rating and shows high correlation and mutual information with the overall rating. It is evident from the plot that the most important trait in an interview is to stay focused. This trait shows a 73 percent correlation with the

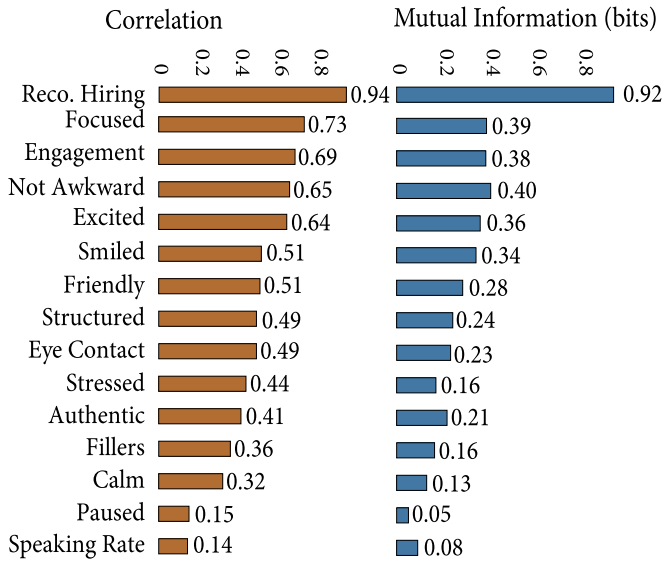


Fig. 5. Correlation and Mutual information between overall rating and ratings on other traits.

overall rating. Some other top traits include possessing an engaging tone, not appearing awkward, being excited, and displaying an appropriate smile. The mutual information and correlation coefficient closely follow the patterns. This plot gives us an insight into what constitutes a good interview.

5.1.3 First (and Last) Impression Matters

We would like to understand how the performance in different interview questions during an interview affects the overall rating. To understand this temporal relationship, we calculated the correlation and mutual information between the ratings for each individual interview question and the ratings for the entire videos. In Fig. 6, we plot this relationship. It is evident from Fig. 6a that performance on the first question correlates most with the overall performance. After the first question, the correlation gradually decays. We can interpret this result as follows: If an interviewee performs well for the first question, it is more likely that he/she will end up receiving an above average rating. It is true in the opposite case as well; if an interviewee performs poorly in the first question, he/she is more likely to receive a poor overall rating. This finding is also supported by existing evidence from psychological point of view [5], [47].

A similar pattern of *first impression matters* holds for ratings on various other traits of the interviewee's behavior, such as whether he/she was excited, smiled, maintained eye contact, talked in engaging tone, or even appeared friendly. Fig. 6b illustrates this. We notice from this figure that there is a sudden spike in correlation for the last question. This indicates the fact that, although the first question matters the most, the interviewee can significantly change the interviewer's perception during the response to the final question.

Fig. 6c shows some traits (e.g., pause, calmness, stress) do not follow the pattern discussed above. However, they have very low correlation values to begin with. We believe it is difficult for Mechanical Turk workers to accurately judge these traits as these judgments demand considerable concentration.

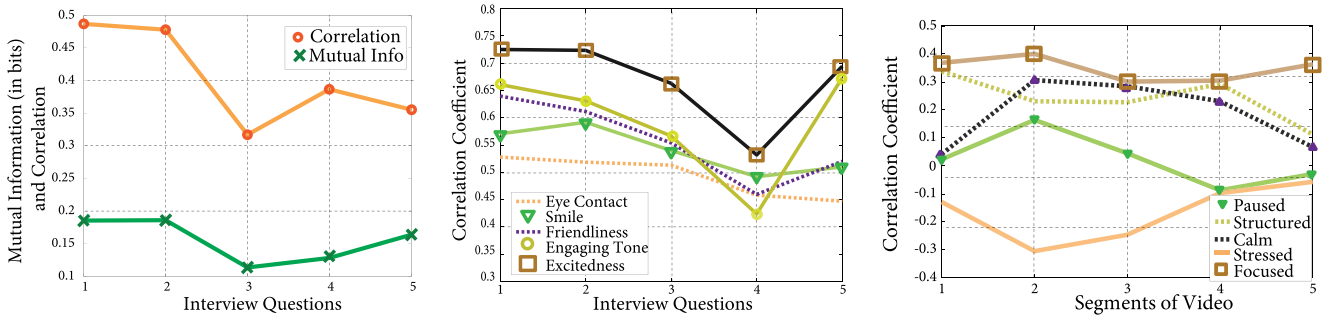
We need to be cautious while interpreting this result. Although the ratings for the first question had maximum correlation with the overall ratings for the entire interview, we can not say whether it is due to the temporal order or the importance of the question itself. However, we would like to emphasize that our mock interviews start with a question about interviewee's background, which is consistent with many real-world job interviews. We also notice that the ratings for individual questions strongly correlate with the ratings obtained for entire videos. Similar observation has also been reported in [48], which shows that predicted ratings from "thin slices" of job interviews are often quite similar to the ratings for entire interviews.

5.2 Prediction Using Automated Features

5.2.1 Prediction Accuracy Using Trained Models

Given the feature vectors associated with each interview video, we would like to provide feedback to users about their overall performance in the interview, the likelihood of getting an offer, and insights into other personality traits that are relevant for job interviews. We train regression models for predicting ratings for a total of 16 traits or rating categories (as shown in Table 1).

The entire dataset has a total of 138 interview videos (for the 69 participants, 2 interviews for each participant). We used 80 percent of the videos for training, and the remaining 20 percent for testing. To avoid any artifacts related to how we split the data, we performed 1,000 random trials. In each trial, we randomly select 80 percent of the participants and include both of their interview videos in the training set, and use the rest for testing. Such participant-level



(a) Correlation and Mutual Information for overall performance (b) Traits following patterns similar to the overall performance (c) Traits not following patterns

Fig. 6. Correlation between ratings of different segments and the rating on the whole interview.

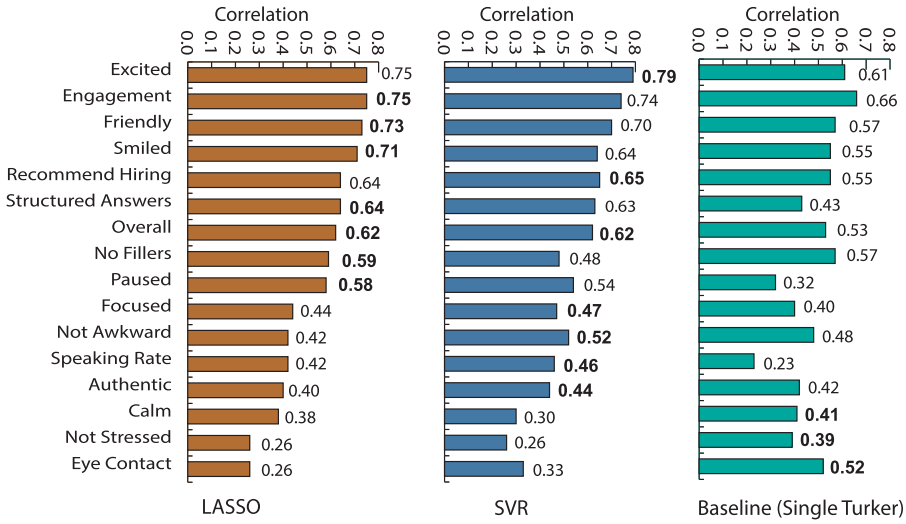


Fig. 7. Regression coefficients using two different methods: Lasso and Support Vector Regression (SVR). We compare our regression models with a baseline method that randomly chooses one of the Turkers' ratings as the predicted rating.

stratified sampling ensures that both videos of a participant end up together either in the training or the test set. We report our results averaged over these 1,000 independent trials. In each trial, we trained 16 different regression models for all 16 traits. For each of the traits, we used exactly the same set of features. The model automatically learned the weights for individual features for each trait.

We measure prediction accuracy by the correlation coefficients between the true ratings and predicted ratings in the test set. Fig. 7 displays the correlation coefficients for different traits, both with SVR and Lasso. The traits are shown in the order of their correlation coefficients. We observe that we can predict several traits with 0.70 or higher correlation coefficients: engagement, excitement, and friendliness. Furthermore, we performed well in predicting overall performance and hiring recommendation scores ($r > 0.62$ for SVR), which are the two most important scores for interview decision. We compare our prediction accuracy with a baseline method based on a single Turker's rating. For each test video, the baseline method randomly selects one of the 9 Turkers, and uses that rating as the predicted score. The Single Turker baseline is a strong one, because (1) it is provided by a human and (2) it is included in the weighted average for ground truth estimation. However, our automatically trained regression models significantly outperform this strong baseline on most of the traits (Fig. 7). We believe it is because our models are trained using a weighted average of 9 ratings, which is more robust and less affected by personal subjective bias.

We also evaluate the learned regression models for a two-class classification task. For each trait, we split the interviews into two groups by the median value for that 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 then vary the threshold on the predicted scores by our regression models in the range $[1, 7]$, and estimate the area under the Receiver Operator Curve (ROC). The random chance baseline area under the curve (AUC) value is 0.50, as we split the classes by the median value. The AUC values for the learned models are presented in Table 5. Again, we observe high accuracies for engagement, excitement, friendliness, hiring recommendation, and the overall score ($AUC > 0.77$ for SVR).

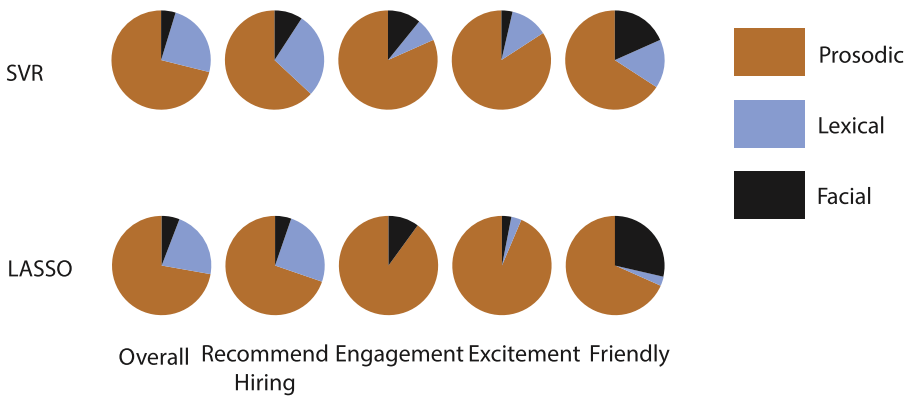
When we examine the traits with lower prediction accuracy, we observe: (1) either we have low interrater agreement for these traits, which indicates unreliable ground truth data (e.g., calm, stressed, structured answer, pause, etc.), or (2) we lack key features necessary to predict these traits (e.g.,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

5.2.2 Feature Analysis

The relative weights of individual features in our regression model can provide valuable insights on essential constituents of a job interview. To analyze this, we observed the features with highest weights for the SVR and the Lasso model. We considered five traits with high accuracy: overall score, recommend hiring, excitement, engagement, and friendliness. We considered the top twenty features in the order of descending weight magnitude, and estimate the

TABLE 5
The Average Area Under the ROC Curve Using SVR, Lasso, and the Single Turker Baseline Method

Trait	SVR	Lasso	Baseline
Excited	**0.91**	0.88	0.76
Engagement	0.84	**0.85**	0.75
Smiled	0.84	**0.86**	0.66
Friendly	**0.81**	0.80	0.71
Recommend Hiring	**0.80**	0.78	0.73
Structured Answers	0.80	**0.82**	0.64
Overall	**0.77**	0.76	0.73
Not Awkward	**0.77**	0.73	0.66
Focused	**0.77**	0.69	0.61
Paused	0.74	**0.75**	0.59
No Fillers	0.73	**0.82**	0.64
Eye Contact	0.68	0.62	**0.71**
Authentic	**0.66**	0.64	0.64
Speaking Rate	**0.63**	0.55	0.56
Calm	0.60	**0.64**	0.62
Not Stressed	0.57	0.57	**0.62**



(a) Relative proportion of the top twenty prosodic, lexical and facial (smile) features as learned by SVR and LASSO classifiers. The weights semantically match our perceptions on the traits

(b) Correlation Coefficients for SVR, for different combinations of facial (F), prosodic (P), and lexical (L) features

Fig. 8. Analysis of relative importance of facial, prosodic, and lexical features.

summation of the weight magnitudes of the features in each of the three categories: prosodic, lexical, and facial features. The relative proportion of prosodic, lexical and facial features are illustrated in Fig. 8a, which shows that both SVR and Lasso assign higher weights to prosodic features while predicting engagement and excitement. This indicates that engagement and excitement are expressed through prosodic features, which agrees with our intuition. 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 (Fig. 5). Since we had smaller number of facial features, the relative weights for facial features is much lower. However, facial features, particularly the smile, were found significant for predicting friendliness. This result provides a solid ground for claiming that smile is very important in order to appear friendly.

Fig. 8b shows the importance of using multimodal features for predicting social traits in job interviews. In most cases, the best correlation coefficient was obtained when we incorporated all three modalities. Although lexical features were critical for predicting overall ratings and likelihood of getting hired, they were not strong predictors of excitement, engagement, and friendliness. Prosodic features played important role for predicting all the five traits, indicating that our speaking style plays a critical role in job interviews.

5.2.3 Recommendation from Our Framework

To better understand the recommended behavior in job interviews, we analyze the feature weights in our regression model. The weights with positive signs and higher magnitudes can potentially indicate elements of a successful job interview. The negative weights, on the other hand, indicates behaviors we should avoid.

We sort the features by the magnitude of their weights and list the top twenty features (excluding the topic features) in Table 6. We see from this table that people having higher speaking rate (higher words per second (*wpsec*), total number of words (*wc*), and total number of unique words (*uc*), etc.) are perceived as better candidates in a job interview. 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* (e.g., they, they'll, them, etc.) and lower usage of words related to *I*. The overall interview performance and likelihood of hiring correlate positively with proportion of positive words, and negatively with proportions of negative words, which agrees with our experience. Individuals who

TABLE 6
Feature Analysis Using the SVR Model

Overall		RecommendHiring		Excited		EngagingTone		Friendly	
avgBand1	-0.12	wpsec	0.139	avgBand1	-0.159	intensityMax	0.174	smile	0.238
wpsec	0.11	avgBand1	-0.132	diffIntMaxMin	0.132	avgBand1	-0.171	mean pitch	0.136
upsec	0.093	Fillers	-0.13	intensityMax	0.124	diffIntMaxMin	0.151	f3STD	-0.106
avgDurPause	-0.09	percentUnvoiced	-0.111	wpsec	0.123	intensityMean	0.146	LipCDt	0.095
Fillers	-0.086	upsec	0.098	smile	0.122	wpsec	0.135	intensityMax	0.094
Quantifiers	0.086	avgDurPause	-0.094	f3STD	-0.115	avgBand2	-0.119	intensityMean	0.09
maxDurPause	-0.076	smile	0.093	intensityMean	0.115	Quantifiers	0.109	diffIntMaxMin	0.089
percentUnvoiced	-0.076	PercentBreaks	-0.09	mean pitch	0.113	f1STD	-0.103	wpsec	0.089
smile	0.074	intensityMean	0.086	nod	0.107	upsec	0.097	f1STD	-0.087
f3meanf1	0.073	f1STD	-0.082	percentUnvoiced	-0.105	f2STDf1	0.096	I	-0.08
f1STD	-0.071	f3meanf1	0.079	intensitySD	0.103	f3meanf1	0.095	Adverbs	0.079
Prepositions	0.07	Quantifiers	0.077	f1STD	-0.101	intensitySD	0.093	fmean3	0.075
Relativity	0.068	Positive emotion	-0.075	PercentBreaks	-0.1	f3STD	-0.091	shimmer	-0.073
f2STDf1	0.067	maxDurPause	-0.074	f2STDf1	0.089	smile	0.082	upsec	0.073
intensityMean	0.065	Prepositions	0.073	wc	0.082	percentUnvoiced	-0.078	avgBand1	-0.07
PercentBreaks	-0.064	nod	0.073	avgBand2	-0.082	f2meanf1	0.077	percentUnvoiced	-0.069
uc	0.061	Articles	0.071	Adverbs	0.082	Cognitive	0.077	PercentBreaks	-0.068
Positive emotion	-0.06	wc	0.07	f3meanf1	0.079	PercentBreaks	-0.076	We	0.063
f2meanf1	0.059	uc	0.069	upsec	0.077	I	-0.071	Sadness	0.062
f3STD	-0.058	Sadness	0.069	Quantifiers	0.075	max pitch	0.07	intensitySD	0.061

We are listing the top twenty features ordered by their weight magnitude. We have excluded the topic features for the ease of interpretation.

smiled more performed better 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 in interview. As we saw earlier, features related to prosody and speaking style are more important to appear excited and engaged. Particularly the amplitude, variations in the voice intensity, and the first 3 formants had high positive weights in our prediction model. Finally, besides smiling, people who spoke more words related to "*We*" than "*I*" were perceived as friendlier.

6 Discussion and Conclusion

We present an automated prediction framework for quantifying social skills for job interviews. The proposed model shows encouraging results and predicts human interview ratings with correlation $r > 0.65$ and AUC ~ 0.80 (compared to the baseline AUC $= 0.50$). Several traits such as engagement, excitement, and friendliness were predicted with even higher accuracy ($r \sim 0.75$, AUC > 0.85). One of our immediate next steps will be to integrate the proposed prediction module with existing automated conversational systems such as MACH to allow valuable real-time feedback to the users.

To our knowledge, the interview dataset used in our experiments is the largest collection of job interview videos, collected under reasonably realistic settings. The interviews are conducted by professional career counselors. We included the questions that would be relevant in most real-world job interviews. Despite efforts to record interviews in realistic settings, we do need to acknowledge several caveats and trade-offs.

All the participants in our dataset were MIT undergraduates, all of junior status, which may introduce a selection bias in our data. In future, we plan to conduct a more comprehensive study over a more general and diverse population. We deliberately chose not to specify a job description to encourage a larger number of student participants. At the time of the study, there were nearly 1,000 junior students present at MIT, and nearly 30 percent were international students. Out of the remaining 700 native English speaking juniors, we were able to recruit 90, which would have been difficult if we had limited our study to a specific job description. However, in the absence of a specific job description, the ground truth ratings may not necessarily correspond to actual hiring decisions, and may show a stronger bias towards non-verbal cues, as there are no specific skill requirements. Furthermore, our mock interviews may lack the stress present in a real job interview. Although we promised to forward the resumes of the top 5 percent candidates to several sponsor organizations, the incentive was not as strong as an actual job offer. In the future, we would like to conduct more controlled experiments with a specific job description and with stronger incentives to induce stress and competition.

We aimed to rate each video with multiple independent judges to avoid personal bias. As a first step, we recruited Turkers as this was scalable, quick, and less expensive. To ensure reliable ground truth ratings, each video was rated using 9 Mechanical Turk workers, and aggregated using the EM algorithm taking the reliability of each worker into account. However, Turkers' ratings may not correspond to professional experts. In future, we plan to collect ratings from a panel of experts, and re-validate the results.

Interestingly, while training regression models using SVR, we obtained better prediction accuracy using the linear kernel, compared to other non-linear kernels (e.g., quadratic, cubic, or Gaussian kernels). This may indicate that our features do not exhibit complicated non-linear interactions. However, the features used in the current models were mostly aggregated features, averaged over the entire duration of the video (e.g., average pitch, average smile intensity). It is plausible that our smile and intonation while uttering a specific word can be a determinant of the final

interview decision. The current aggregated features are incapable of modeling such temporal interactions. Modeling fine-grained temporal features across multiple modalities is left as future work. We would also like to group the words in interviewees' transcripts using standard clustering algorithms (e.g., Brown Clustering, Vector Representations based Clustering), and use the relative frequencies of clusters as additional lexical features. Our lexical features were extracted from manual speech transcripts generated by Turkers. However, these lexical features could potentially be extracted using automated speech recognition and filler word detection systems. Recent Long-Short-Term-Memory Recurrent Neural Network (LSTM-RNN) models for speech recognition [14] and disfluency detection [49] achieve ~ 85 percent or even higher accuracy, which is not much worse than Turkers' captioning quality.

Our existing system is trained to predict ratings based on entire interview videos, and one has to wait till the end of the interview in order to receive feedback. Real-time feedback has been shown to be helpful for different behavioral domains, for example improving public speaking skills [50]. Extending our prediction framework to provide real-time feedback can be an exciting future direction.

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 in job interviews can provide valuable insight and understanding regarding job interviews and human behavior in general.

Caveats aside, the results presented in this article show the importance of including multiple modalities while analyzing our social interactions. The analysis of the feature weights learned by our prediction models provides quantitative insights to the determinants of successful job interviews. With the knowledge presented in this article, 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 [51], train customer service professionals, or even help medical professionals with telemedicine.

Acknowledgments

This work was partially supported by Google Faculty Research Award, and Microsoft Azure for Research Award. We would like to acknowledge the MIT undergraduate students who consented to share their interview data to advance science. We would also like to thank Leon Weingard for helping with transcribing the audio, and Michaela Kerem for her extensive feedback. Finally, we are grateful to anonymous reviewers and the associate editor for all the helpful comments and reviews.

References

[1] J. Gottman, H. Markman, and C. Notarius, "The topography of marital conflict: A sequential analysis of verbal and nonverbal behavior," *J. Marriage Family*, vol. 39, pp. 461–477, 1977.

[2] J. M. Gottman, *The Mathematics of Marriage: Dynamic Nonlinear Models*. Cambridge, MA, USA: MIT Press, 2002.

[3] J. A. Hall, D. L. Roter, D. C. Blanch, and R. M. Frankel, "Nonverbal sensitivity in medical students: Implications for clinical interactions," *J. General Internal Med.*, vol. 24, no. 11, pp. 1217–1222, 2009.

[4] N. Ambady and R. Rosenthal, "Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness," *J. Personality Social Psychology*, vol. 64, no. 3, 1993, Art. no. 431.

[5] J. R. Curhan and A. Pentland, "Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes." *J. Appl. Psychology*, vol. 92, no. 3, 2007, Art. no. 802.

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

[7] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *Affective Computing and Intelligent Interaction*. Berlin, Germany: Springer, 2007, pp. 71–82.

[8] V. Soman and A. Madan, "Social signaling: Predicting the outcome of job interviews from vocal tone and prosody," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.*, 2010.

[9] 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." *J. Appl. Psychology*, vol. 86, no. 5, 2001, Art. no. 897.

[10] R. A. Posthuma, F. P. Morgeson, and M. A. Campion, "Beyond employment interview validity: A comprehensive narrative review of recent research and trends over time," *Personnel Psychology*, vol. 55, no. 1, pp. 1–81, 2002.

[11] T. Macan, "The employment interview: A review of current studies and directions for future research," *Hum. Resource Manag. Rev.*, vol. 19, no. 3, pp. 203–218, 2009.

[12] I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque, "Automated prediction and analysis of job interview performance: The role of what you say and how you say it," in *Proc. 11th IEEE Int. Conf. Workshops Automat. Face Gesture Recognit.*, 2015, pp. 1–6.

[13] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, 2005, pp. 677–682.

[14] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1468–1472.

[15] M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard, "MACH: My automated conversation coach," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2013, pp. 697–706.

[16] M. E. Hoque, D. J. McDuff, and R. W. Picard, "Exploring temporal patterns in classifying frustrated and delighted smiles," *IEEE Trans. Affective Comput.*, vol. 3, no. 3, pp. 323–334, Jul.-Sep. 2012.

[17] E. Sapir, *Selected Writings of Edward Sapir in Language, Culture and Personality*, vol. 342. Berkeley, CA, USA: Univ. California Press, 1985.

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

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

[20] M. I. Tanveer, R. Zhao, K. Chen, Z. Tiet, and M. E. Hoque, "Automanner: An automated interface for making public speakers aware of their mannerisms," in *Proc. 21st Int. Conf. Intell. User Interfaces*, 2016, pp. 385–396.

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

[22] R. Ranganath, D. Jurafsky, and D. McFarland, "It's not you, it's me: Detecting flirting and its misperception in speed-dates," in *Proc. Conf. Empirical Methods Natural Language Process.*: vol. 1, pp. 334–342, 2009.

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

[24] A. Stark, I. Shafran, and J. Kaye, "Inferring social nature of conversations from words: Experiments on a corpus of everyday telephone conversations," *Comput. Speech Language*, vol. 28, no. 1, pp. 224–239, 2014.

[25] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal recognition of personality traits in social interactions," in *Proc. 10th Int. Conf. Multimodal Interfaces*, 2008, pp. 53–60.

[26] D. Sanchez-Cortes, J.-I. Biel, S. Kumano, J. Yamato, K. Otsuka, and D. Gatica-Perez, "Inferring mood in ubiquitous conversational video," in *Proc. 12th Int. Conf. Mobile Ubiquitous Multimedia*, 2013, pp. 22:1–22:9.
[27] J.-I. Biel, V. Tsiminaki, J. Dines, and D. Gatica-Perez, "Hi Youtube!: Personality impressions and verbal content in social video," in *Proc. 15th ACM Int. Conf. Multimodal Interaction*, 2013, pp. 119–126.
[28] D. Sanchez-Cortes, S. Kumano, K. Otsuka, and D. Gatica-Perez, "In the mood for vlog: Multimodal inference in conversational social video," *ACM Trans. Interactive Intell. Syst.*, vol. 5, no. 2, 2015, Art. no. 9.
[29] L. Teijeiro-Mosquera, J.-I. Biel, J. L. Alba-Castro, and D. Gatica-Perez, "What your face vlogs about: Expressions of emotion and big-five traits impressions in Youtube," *IEEE Trans. Affective Comput.*, vol. 6, no. 2, pp. 193–205, Apr.-Jun. 2015.
[30] L. S. Nguyen and D. Gatica-Perez, "Hirability in the wild: Analysis of online conversational video resumes," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1422–1437, Jul. 2016.
[31] L. Nguyen, D. Frauendorfer, M. Mast, and D. Gatica-Perez, "Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1018–1031, Jun. 2014.
[32] K. Anderson, et al., "The TARDIS framework: Intelligent virtual agents for social coaching in job interviews," in *Advances in Computer Entertainment*. Berlin, Germany: Springer, 2013, pp. 476–491.
[33] T. Baur, I. Damian, F. Lingenfelser, J. Wagner, and E. André, "Nova: Automated analysis of nonverbal signals in social interactions," in *Human Behavior Understanding*. Berlin, Germany: Springer, 2013, pp. 160–171.
[34] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (version 5.1. 05)[computer program]. retrieved May 1, 2009," Phonetic Sciences, University of Amsterdam, Spuistraat, The Netherlands, 2009.
[35] J. W. Pennebaker, M. E. Francis, and R. J. Booth, *Linguistic Inquiry and Word Count: LIWC 2001*, Mahwah: Lawrence Erlbaum Associates, vol. 71, 2001.
[36] R. Caruana and D. Freitag, "Greedy attribute selection." in *Proc. Int. Conf. Mach. Learning*, 1994, pp. 28–36.
[37] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learning Res.*, vol. 3, pp. 993–1022, 2003.
[38] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[39] B. Froba and A. Ernst, "Face detection with the modified census transform," in *Proc. 6th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2004, pp. 91–96.
[40] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 1034–1041.
[41] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *Int. J. Comput. Vis.*, vol. 91, no. 2, pp. 200–215, 2011.
[42] V. C. Raykar "Learning from crowds," *J. Mach. Learning Res.*, vol. 99, pp. 1297–1322, 2010.
[43] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, 2004.
[44] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statistical Soc. Series B (Methodological)*, vol. 58, pp. 267–288, 1996.
[45] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Edu. Psychological Meas.*, vol. 30, no. 1, pp. 61–70, 1970.
[46] J. L. Fleiss, B. Levin, and M. C. Paik, "The measurement of interrater agreement," *Statistical Methods Rates Proportions*, vol. 2, pp. 212–236, 1981.
[47] T. W. Dougherty, D. B. Turban, and J. C. Callender, "Confirming first impressions in the employment interview: A field study of interviewer behavior." *J. Appl. Psychology*, vol. 79, no. 5, 1994, Art. no. 659.
[48] L. S. Nguyen and D. Gatica-Perez, "I would hire you in a minute: Thin slices of nonverbal behavior in job interviews," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2015, pp. 51–58.
[49] V. Zayats, M. Ostendorf, and H. Hajishirzi, "Disfluency detection using a bidirectional LSTM," Interspeech 2016, pp. 2523–2527, 2016, doi: 10.21437/Interspeech.2016-1247.
[50] M. I. Tanveer, E. Lin, and M. E. Hoque, "Rhema: A real-time in-situ intelligent interface to help people with public speaking," in *Proc. 20th Int. Conf. Intell. User Interfaces*, 2015, pp. 286–295.
[51] M. J. Smith, "Brief report: Vocational outcomes for young adults with autism spectrum disorders at six months after virtual reality job interview training," *J. Autism Developmental Disorders*, vol. 45, no. 10, pp. 3364–3369, 2015.

Iftekhar Naim received the BSc degree in computer science from Bangladesh University of Engineering and Technology (BUET), in 2007, and the MSc degree in electrical and computer engineering from the University of Rochester, in 2011. He is currently working toward the PhD degree in the Computer Science Department, University of Rochester. He is working on natural language processing and multimodal data analysis, under the supervision of Prof. Daniel Gildea. He did several research internships with Bosch Research and Google Inc. He is a student member of the IEEE, the AAAI, and the ACL.

Md. Iftekhar Tanveer received the BSc degree in electrical and electronic engineering (EEE) from Khulna University of Engineering and Technology (KUET), Bangladesh, in 2007, and the MSc degree in electrical and computer engineering (ECE) from the University of Memphis, Tennessee, in 2011. He is currently working toward the PhD degree in the Department of ECE, University of Rochester. He works as a research assistant in the Department of Computer Science, ROC-HCI lab under supervision of Prof. Mohammed Ehsan Hoque. He did an internship with Ebay Inc. in 2012. He is currently interested in machine learning research for applications in human behavioral analysis. He is a student member of the IEEE and the ACM.

Daniel Gildea received the PhD in computer science from the University of California, Berkeley, in 2001. He is an associate professor in the Department of Computer Science, University of Rochester. His research investigates various aspects of natural language processing, including machine translation, language understanding, analysis of multimodal data, and the computational complexity of parsing and translation problems.

Mohammed Ehsan Hoque received the PhD degree from the Massachusetts Institute of Technology, in 2013. He is an assistant professor of computer science with the University of Rochester where he leads the ROC HCI Group. He has previously held positions with Goldman Sachs, Walt Disney Imagineering Research & Development, and IBM T. J. Watson Research Center. Hoque's research efforts are about developing computational approaches to decipher and model the unwritten rules of human communication as well as designing and deploying new interactive systems. His work has received Best Paper Award at ACM Ubiquitous Computing (UbiComp 2013), Best Paper Nominations at IEEE Automated Face and Gesture Recognition (FG 2011), and ACM Intelligent Virtual Agents (IVA 2006). His research has been recognized with Google Faculty Research Award and MIT TR35 Award. He is a member of the IEEE.