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