Analysis of Dyadic Interaction in an Interview Setting

Suresh Alse, Bhavishya Sharma, Jay Priyadarshi, Abhishek Sharma

November 19, 2016

1 Introduction

Interviews are often hard to be judged. It is often left in the hands of the interviewer(s) to measure the hirability of the candidates. This is fundamentally flawed as this heavily relies of interviewers' mood and personality. Also, in most cases multiple interviewers interview for the same roles which makes this process even less scientific as it is almost impossible to fairly aggregate the opinions of interviewers.

There has been tons of research by psychologists and career experts about what one should do in order to succeed in an interview (Huffcutt, Conway, Roth, & Stone, 2001). From this, we know that things like smiling, using a confident tone and making good eye contact can contribute a lot in an interview. However, these observations are often based on intuition and experience. Hence, It is hard to automate and quantify hirability of candidates. Also, there is a common misconception that content of the interviewee's responses is the sole determinant of the job interview. However, it is seen that non verbal aspects are as important if not more important than verbal responses (Mehrabian et al., 1971).

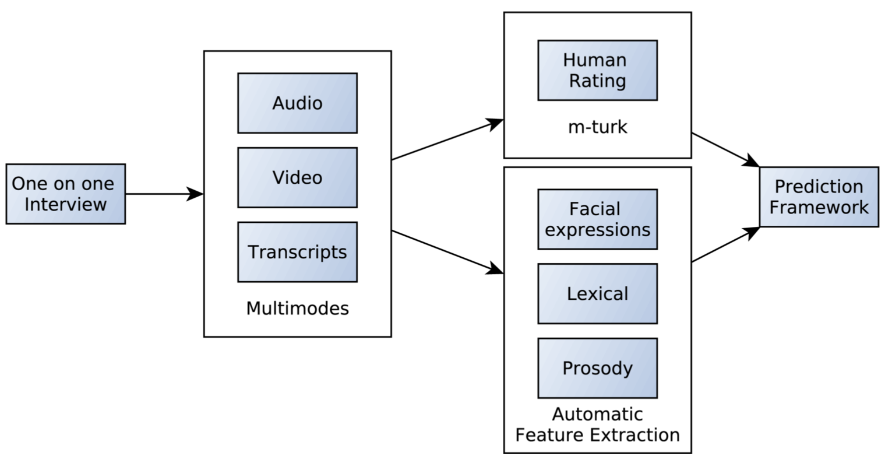

Figure 1: Proposed Framework

In this project we would like to build a computational framework using which interviewers and interviewees can use it to analyze interviews and obtain the following.

- Automatically predict the overall score of the interview.
- Quantify verbal and nonverbal behavior of the interviewee towards the success in the interview.

- Automatically recommend aspects to be improved for better overall score.
- Timeline that shows how well the interview progressed with respect to time.

In order to achieve this we propose a framework as shown in the figure below. We use a one on one interview data comprised of three modes (audio, video and textual). Then, we extract multimodal features (facial expressions, lexical and prosody) and predict the overall score of the interview, how likely the candidate is going to be hired and other traits required for the interview process.

2 Related Work

A lot of the research in the field of mulitmodal analysis of interaction has focused on speech and visual analysis of data. For instance, in Rough'n'Ready: A Meeting Recorder and Browser (Kubala, Colbath, Liu, & Makhoul, 1999), they provide a way to recognize speech in the form of a BBN Byblos Speech Recognition System, where they also provide a mechanism to browse and retrieve speech data with the help of a speech index. Speaker identification is also described in The Meeting Project at ICSI (Morgan et al., 2001), where the acoustic model consisted of gender-dependent, bottom-up clustered (genonic) Gaussian mixtures. Further, leveraging speech recognition, topic detection in a meeting room scenario is described in Advances in Automatic Meeting Record Creation and Access, where they use a variant of Hearst's TextTiling algorithm in order to automaticaly segment the transcript into topically coherent passages.

As far as visual analysis is concerned, we can find examples of that in SMaRT: The Smart Meeting Room Task at ISL (Waibel et al., 2003), where they provide a mechanism to track people and identify them as they move around a Meeting Room using multiple cameras and advanced computer vision techniques. Another good example of that would be Distributed Meetings: A Meeting Capture and Broadcasting System (Cutler et al., 2002) where they augment the meeting room for remote viewers by adding cameras and other functionalities.

A major focus on such speech and visual processing (as provided above) has been focused on individuals, however, even when the researchers examine a meeting space. Our aim is to analyze dyadic communication where we don't just monitor an individual, but we attempt to find multimodal cues (such as back-channels among others) which would then uncover the underlying mechanism of a job interview.

There has been research on analyzing behavior of a group as compared to an individual, as is exemplified by research like The KidsRoom: A Perceptually-Based Interactive (Bobick et al., 1999) and Immersive Story Environment and A Bayesian Computer Vision System for Modeling Human Interactions (Oliver, Rosario, & Pentland, 2000). However, the research here focuses on problem specific "primitive tasks", and therefore involves a much more constrained examination, which is in a sharp contrast to a sort of free-flowing, spontaneous (dyadic) interaction that we would have hoped for.

While our system focuses on some form of speech and visual processing, and also incorporates analysis of dyadic interaction as a whole, we provide a way to analyze the interaction in a much more unconstrained manner, identifying key multimodal cues, unraveling the underlying operating factors of a job interview by treating an interview as "more than a some of its parts" and hopefully, to come up with capabilities to automatically predict the overall score of an interview, quantify verbal and non verbal behavior of the interviewee towards the success in the interview, automatically recommend aspects to be improved for a better overall score, and a timeline to show how well an interview progressed with respect to time.

3 Dataset

We use the MIT Interview Dataset (Naim, Tanveer, Gildea, & Hoque, 2015) for this project that we obtained by contacting the authors of the project. It consists of 138 recordings of mock interviews of students from MIT, seeking internships. The interviews were conducted in a one on one interview fashion. Both interviewers and interviewees were equipped with microphones which allows us to extract and differentiate between the speakers easily. Cameras were used to capture the video of the interviewee during the process as shown in the Figure 2. The interviews were conducted by two professional career counselors with over five years of interviewing experience. All participants are native english speakers (this is very important because in our approach things like confidence, fluency, etc are considered). For every participant, two rounds were conducted - before and after intervention. Overall, 69 students permitted the use of recordings for research purposes. Hence we have a total of 138 recordings of lengths between 3 minutes to 8 minutes (average: 4.7 minutes per interview). Every interview consisted of interviewer asking the interviewee, five questions and no job description was given to the interviewees. The researchers who collected this data claim that this is the largest collection of job interview videos conductedby professionals.

To rate the interview, Amazon mechanical turk workers were used. Each turker watched the interview videos and rated the interviews by answering 16 assessment questions 1 on seven point scale. Questions about "Overall rating" and "Recommend Hiring" captures overall score where as other questions capture higher level behavior.

Engagement
Excited
Friendly
Smiled
NoFillers
RecommendedHiring
Overall
EyeContact
NotAwkward
StructuredAnswers
Calm
Focused
NotStressed
Authentic
Paused
SpeakingRate

Table 1: Assessment questions

The dataset also consists of transcripts of all the interviews. This was made possible by Amazon mechanical turk workers hired by the researchers. Also, they were instructed to include filler words such as "like", "uh", "umm" along with cues like "[long pause]", "[smiling]" etc which are very useful for our process.

We also tried semaine-db (McKeown, Valstar, Cowie, Pantic, & Schroder, 2012) which seemed good for this project. However, it just consisted data of two individuals talking to each other and was in no way an interview setting. We also considered using AMI database which consisted of a group discussing about a particular topic for a day. However, this had additional problems such as multiple people in a frame, etc and moreover similar to semaine-db, this was not a interview setting. Also, as this needs a considerable amount of data in a given setting and then requires amazon mechanical turkers, creating our own dataset seemed farfetched. Hence, we chose MIT Interview Dataset which is perfect for our project.

Figure 2: One of the 138 interviews.

3.1 Inter-rater agreement

Jay TODO

3.2 Feature analysis

We also did some analysis on the features extracted to get more insights to do feature selection while doing regression. For every feature we try to find the correlation between the features and the scores of assessment questions.

3.2.1 Prosodic Features

The extracted prosodic features consists of energy, power, pitch etc. Every interview is divided into four segments corresponding to four different questions. After averaging out all the features for each of the segments, we try to match it with the scores assigned by the turkers for each of the assessment questions. We draw a scatter plot and try to fit a line to see how relevant the score is to each of these features. A positive slope indicates that with the increase in the value of the feature, a higher score would be assigned. A negative slope would indicate that with the increase in the value of the feature, a lower score would be assigned. A near zero slope would indicate that this feature wouldn't matter and we can neglect it in regression. Figures 3, 4, 5 and 6 are some of the 1026 graphs that were drawn to visualize this. In Figure 3 we can see that with the increase in energy, interviewees are likely to be more excited. Figure 4 indicates that the recommended rating score drops with the increase in shimmer. Figure 6 indicates that Min pitch and Engaged score are not related and hence we can ignore it.

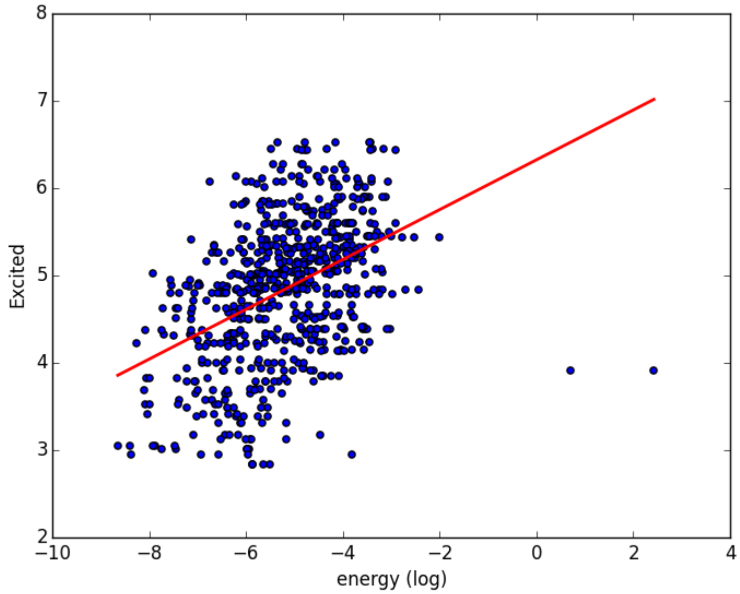

Figure 3: Energy vs Excited

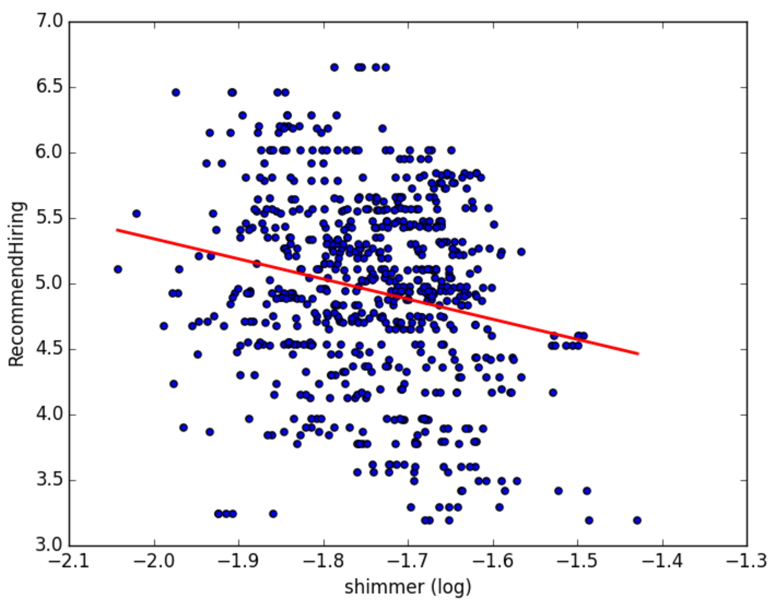

Figure 4: Shimmer vs Recommended Hiring

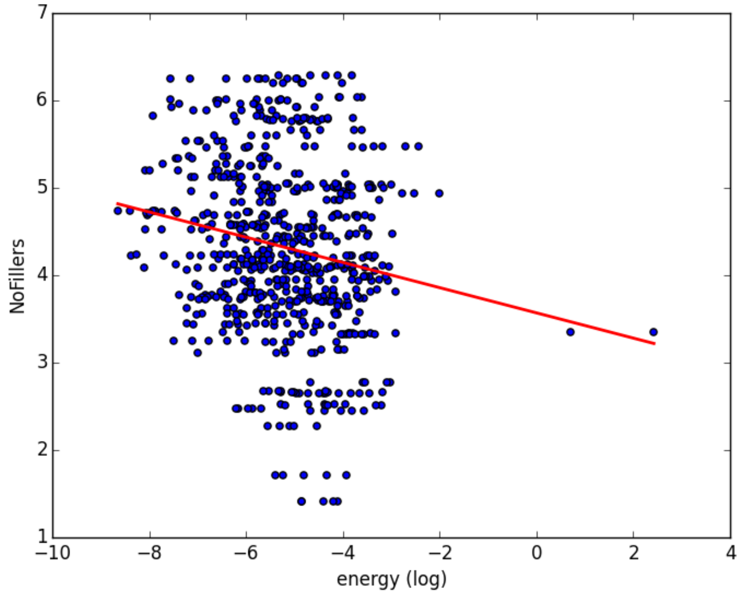

Figure 5: Energy vs No. of Fillers

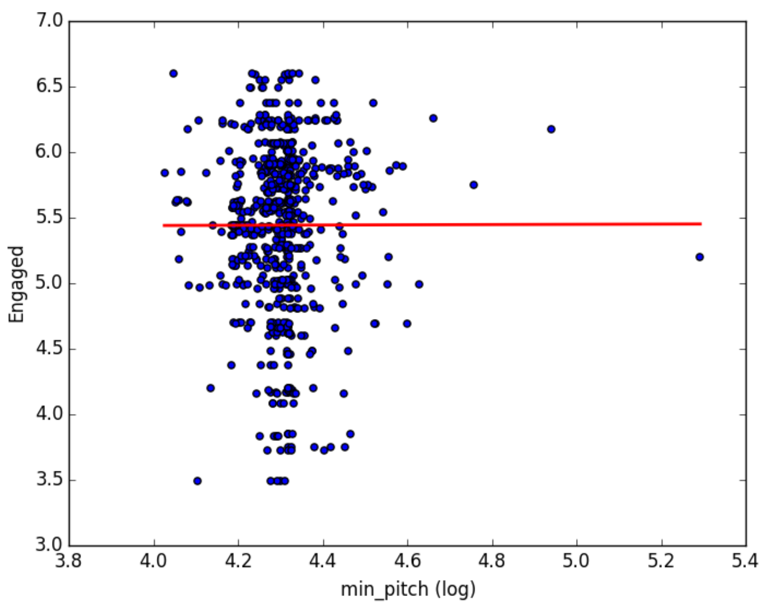

Figure 6: Min Pitch vs Engaged

4 Methodology

This section we describe the overall approach towards building the proposed framework.

4.1 Feature Extraction

We consider three categories of features in our approach i.e Prosodic features, lexical features and facial features. Also, as the data provides with necessary transcripts from m-turkers along with filler words, we don't have to use any automatic speech recognition. Hence lexical features can be extracted directly from transcripts.

4.1.1 Prosodic Features

In order to extract prosodic features from the audio, we are planning to use an open source speech analysis tool called PRAAT (Naim et al., 2015). We obtain the features from parts where interviewer and interviewee speak separately. We then obtain averages over every response and over five responses. Thus we can both reduce dimensionality and retain temporal nature of the prosody that is required for our analysis.

According to some of the previous research (Frick, 1985), pitch, intensity, characters of first three formants and spectral energy are found to be more representative of our behavior. For every feature we extracted mean, variance, minimum and maximum values. We also extracted additional features such as pauses, non-uniform pitch and intensity of speeches as it will help in determining overall score of the interview.

4.1.2 Lexical Features

Word count is often used as lexical feature in many applications. However, we only have limited data; hence, we will not be able to use it as it would result in sparse high dimentional feature vectors. To resolve this problem, we will use Latent Dirichlet Allocation (LDA) to learn 20 topics from interview dataset. Then, we use the relative weights of these topics in every interview as lexical features.

Also, we know that speaking rate and fluency can be indicators of a good interview. Hence, we are also planning to use additional features such as words per second, unique words per second, filler word count and unique word count.

4.1.3 Facial Features

Facial features are very important and are hard to be quantified. In this project, we will extract features from every frame in the video sequence. There are many tools such as OKAO, CLM and GAVAM that can give use features such as smile level, position of head so that we can also incorporate features like nodding. We still need to investigate as to which one will be more appropriate to our problem. We are also planning on using OpenFace which uses deep convolutional neural networks. OpenFace will help us with Facial Landmark detection, head pose estimation, Action Unit detection for inferring various distinct facial expressions. It also helps with Eye gaze estimation. Similar to other features, we will calculate the overall averages and averages over chunks of time frames.

After extracting all the features, we will concatenate them to form a feature vectors. We will normalize all the features to have zero mean and unit variance to eliminate bias.

4.2 Score Predictions

We will train regression models using the features extracted to predict the overall score of the interview, extent of contribution of features towards the score and scores of time frames in a given interview. For this we are planning to use either Support Vector Regression or LASSO. We can also use other approaches such as logistic regression (L1 reg) or Gausian Mixture models. We will decide the best algorithm using turkers' rating and how well our approach matches them. In order to find the extent of contribution of features (required for recommendations) we can just estimate weights of the features in determining the given result.

5 Timeline

Mid Term	Prosodic, lexical and facial feature extraction. Aggregating the features to form feature vectors of all required types.
Post Mid Term	Use evaluation techniques for feature extractions. Develop a plug and play like system to test out different approaches of regression.
Final	Use SVC and Lasso to estimate scores of the interviews and find accuracy of our approach. Build a system that can generate reports, graphs and recommendations.

References

Bobick, A. F., Intille, S. S., Davis, J. W., Baird, F., Pinhanez, C. S., Campbell, L. W., ... Wilson, A. (1999). The KidsRoom: A perceptually-based interactive and immersive story environment. *Presence: Teleoperators and Virtual Environments*, *8*(4), 369–393.

Cutler, R., Rui, Y., Gupta, A., Cadiz, J. J., Tashev, I., He, L.-w., ... Silverberg, S. (2002). Distributed meetings: a meeting capture and broadcasting system. In *Proceedings of the tenth acm international conference on multimedia* (pp. 503–512).

Frick, R. W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin*, *97*(3), 412.

Froba, B., & Ernst, A. (2004). Face detection with the modified census transform. In *Automatic face and gesture recognition, 2004. proceedings. sixth ieee international conference on* (pp. 91–96).

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, *86*(5), 897.

Kubala, F., Colbath, S., Liu, D., & Makhoul, J. (1999). Rough'n'Ready: a meeting recorder and browser. *ACM Computing Surveys (CSUR)*, *31*(2es), 7.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroder, M. (2012). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, *3*(1), 5–17.

Mehrabian, A., et al. (1971). *Silent messages* (Vol. 8). Wadsworth Belmont, CA.

Morgan, N., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Janin, A., ... Stolcke, A. (2001). The meeting project at ICSI. In *Proceedings of the first international conference on human language technology research* (pp. 1–7).

Naim, I., Tanveer, M. I., Gildea, D., & Hoque, M. E. (2015). Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *Automatic face and gesture recognition (fg), 2015 11th ieee international conference and workshops on* (Vol. 1, pp. 1–6).

Oliver, N. M., Rosario, B., & Pentland, A. P. (2000). A bayesian computer vision system for modeling human interactions. *IEEE transactions on pattern analysis and machine intelligence*, *22*(8), 831–843.

Waibel, A., Schultz, T., Bett, M., Denecke, M., Malkin, R., Rogina, I., … Yang, J. (2003). SMaRT: The smart meeting room task at ISL. In *Acoustics, speech, and signal processing, 2003. proceedings.(icassp'03). 2003 ieee international conference on* (Vol. 4, pp. IV–752).