Appears: IEEE In'tl Conference on Acoustics, Speech and Signal Processing, Dallas TX March 2009

SOCIAL SIGNALING: PREDICTING THE OUTCOME OF JOB INTERVIEWS FROM VOCAL TONE AND PROSODY

Vikrant Soman
Electrical Engineering Department
University of Wisconsin-Madison
soman@wisc.edu

Anmol Madan
MIT Media Laboratory
Massachusetts Institute of Technology
anmol@media.mit.edu

ABSTRACT

What does it take to succeed in a job interview? Recruiters, career coaches and social psychologists alike have highlighted the role of speaking style, confidence and demeanor in the face-to-face interview process. Past work suggests that non-linguistic verbal cues play an important role in the outcomes of interviews, and that social signaling measures are quite effective in predicting behavioral outcomes in different social interactions (e.g. negotiations, dating). In this paper, we quantify the non-linguistic speaking style of engineering school students in practice job interviews, using features extracted from their vocal tone and prosody. We find that successful candidates have a characteristic speaking style and these vocal features can be used to build a predictive model of the interview outcomes, with over 85% accuracy.

Index Terms— social signaling, job interviews, speech features, social psychology

1. INTRODUCTION

Face-to-face communication in humans is a highly complex process involving verbal content, non-verbal signaling, gestures and body posture. While verbal communication is explicit and is easily interpreted, non-verbal communication is subtle and implicit. Nonetheless, it has been well established that both channels of communication affect conversational dynamics and influence the relationships between individuals [4]. In non verbal communication subtle aspects of speech like tone, intensity, pitch etc. are categorized as *non-verbal paralinguistic communication,* and observed in cases where a person is described as 'driving the conversation' or 'setting the tone' of the conversation [7].

Face-to-face interviews play pivotal role in the hiring process for companies and help the recruiter evaluate the candidate's skills, motivation and personality traits. There are different types of job interviews, including broad or screening interviews (e.g. 'why do you want to work for this company?'), behavioral interviews (e.g. 'give me an example when you managed multiple projects'), technical interviews (related to particular job skills) amongst a few. Screening interviews are usually conducted by Human Resource (HR) managers, and are meant to gauge the overall motivation, attitude and aptitude of the candidate. These interviews are typically less structured and subjective in nature, and the decision of the interviewer is based to a greater extent on the quality of interaction with the candidate. Besides the conversational content, other aspects like demeanor, physical appearance, dressing style have been shown to influence the decision of the interviewer. Research suggests that the motivation, confidence and attitude of the candidate reflected in non-linguistic verbal cues plays a significant role in determining the outcome of an interview [1, 3, 4, 9].

Researchers in applied social psychology have used various models to encode the communication in job interviews and then determine the influence of different cues [9]. In this paper, we quantify non-linguistic communication in job interview conversations using social signaling measures proposed by Pentland [7], which have been successfully used to predict outcomes in interactions like negotiations, business elevator pitches and speed dating [5, 8]. Reliable quantification of social signals can account for almost a third of the variance in behavioral outcome (70 – 85 % binary decision accuracy), and non-linguistic social signals have been found to be as important as linguistic content for certain types of interactions.

In the next section, we describe the non-linguistic measures based on tone and prosody in more detail. In Section 3, we explain the experiment design and data collection. In Section 4, we analyze the characteristic speaking style of successful candidates and build a predictive model of the outcome of the interview based on these features.

2. NON-LINGUISTIC SOCIAL SIGNALS

Pentland proposed functions of prosody and speaking style, which attempt to capture the *social signaling* between individuals. In our work, we compute these four measures of speaking style and use them to predict the performance of individual speakers. These measures are summarized below, and explained in more detail here [7].

To calculate these signaling measures, it is required to

identify the voicing/non-voicing and speaking/non-speaking segments from raw audio. This is done using a linked Hidden Markov Model proposed by Basu [2], wherein energy independent features are used for extracting the voiced segments and then these segments are grouped into speech regions. The following *social signals* are then calculated for each conversation.

Activity Measure:
Activity is defined as the z-scored percentage of fraction of speaking time plus the frequency of voiced segments. In our case we considered both the quantities as separate features. Fraction speaking time was measured as the ratio of duration of speaking frames in the conversation to the total duration of the conversation. Voicing frequency was measured as the rate of voicing segments found in the speaking region of the conversation.

Fraction speaking time = s / n
where s = speaking frames,
n= total frames in speech segment

Voicing rate = v / (v + u)
where v = voiced frames, u = unvoiced frames

Engagement Measure:
Engagement is defined as the z-scored influence each person has on the other's turn-talking. In a two person conversation, individual turn-taking dynamics influence each other and that can be modeled as a Markov process [7]. A measure of the engagement can be obtained by quantifying the influence each person has on the other. Thus speaking states of each individual are modeled using a Hidden Markov Model (HMM) and the measure of the coupling between these two HMMs gives an estimate of the influence each person has on the other which we use as the engagement measure. Specifically, the influence model approach summarizes the directed coupling between two chains to a single alpha parameter.

$$P(S^i_t \mid S^i_{t-1} \ldots\ldots S^N_{t-1}) = \Sigma\, \alpha_{ij}\, P(S^i_t \mid S^j_{t-1})$$

where α_{ij} *= influence coupling parameter between interacting chains i and j*
$P(S^i_t)$ *= prob. that chain i is in state S at time t*
i = chains from 1 to N
t = discrete time steps

Emphasis Measure:
Emphasis is the variation in prosodic emphasis. For measuring it, in each voiced segment we extract the mean energy, frequency of the fundamental format, and the spectral entropy. After averaging over longer periods we get estimates of the mean-scaled standard deviation of the energy, formant frequency and spectral entropy. The z-scored sum of these standard deviations is taken as a measure speaker emphasis. It has been found that such emphasis can be either purposeful (e.g., indicate significance) or unintentional (e.g., physiological stress caused by discomfort) [7].

$$\textit{Emphasis measure} = \sum \textit{std}(\epsilon) + \textit{std}(\mu) + \textit{std}(\rho)$$

where ϵ = formant frequency,
μ = spectral entropy,
ρ = energy in frame ,
and std is standard deviation

Mirroring Measure:
Often in a conversation we observe back-and-forth exchanges typically consisting of single words like ('OK?', 'Yes!', 'done') and short interjections (< 1 sec) like 'uh-huh', 'hummh'. This is termed as mirroring behavior, in which the speaking style of one participant is mirrored by the other, is considered to signal empathy [7]. Thus we found out short voicing segments for both speakers that were in closely spaced in time (< 1sec) indicating a quick utterance exchange. Total number of such exchanges in the conversation was taken as a measure of mirroring

Mirroring measure = {(S1 (i) – S2 (j)) ≤ 1sec } / n
where S1(i) = time of occurrence of short speaking frame(<1s) for speaker 1
S2(j) = time of occurrence of short speaking frame(<1s) for speaker 2
n = length of speech segment in seconds
{ } = total number of such pairs in time n

3. DATA COLLECTION

Data for our analysis was collected from practice job interviews at an engineering undergraduate school. The interviewers were two Human Resource (HR) managers from the graduate school of business affiliated with the same university. The interviews themselves were non-technical in nature, and consisted of typical questions part of an initial HR screening interview. All interviews were consistent in composition and averaged approximately ten minutes in length, and followed the format below:

1. *'Hi. Please introduce yourself. '*
2. *'Please briefly describe your academic background and coursework.'*
3. *'What motivated you to follow this particular field of engineering '*
4. *Example scenario related to work or professional commitment*
6. *'What extra-curricular activities do you participate in?'*
7. *'Why would you be a good fit for our company?'*

After each interview, the interviewers graded the candidates on the following criteria:

1. *Overall rating for the candidate (scale 1-5)*

2. Did the candidate appear confident during the interview? (scale 1-5)
3. Was the candidate sufficiently engaged in the interview? (scale 1-5)
4. Were you impressed by the candidate? (scale 1-5)

Students were briefed that these were practice HR interviews arranged to help them prepare for on-campus recruitment. It was also disclosed that the interviewers were experienced HR managers. This ensured that the interviews were conducted in all their seriousness.

4. ANALYSIS AND RESULTS

For this study, twenty-six candidates within the ages of 20 - 23 (juniors, and seniors) were interviewed. The conversations were checked for sound quality and 5 samples were discarded on account of bad recording quality and excessive hum.

Each interview was split into two chunks, of approximately five minutes each. Training labels for each chunk were based on the responses provided by the interviews for the four training criterion.

4.1 Non-linguistic Speaking Style of Highly-Rated Candidates and their Interviewers

Do successful candidates have a characteristically different speaking style? Based on distribution of the 'overall rating' responses, the candidates can be divided into two classes with approximately equal samples —highly-rated (i.e. overall rating > 3) and poorly-rated (overall rating <= 3).

In our dataset, we find that these highly-rated candidates have, on average, higher activity levels (mean = 0.52 for highly-rated candidates, mean = 0.33 for poorly-rated candidates, $p < 0.0001$). These candidates also show lower levels of vocal emphasis during the interview conversation (mean = 2.25 for highly-rated candidates, mean = 3.25 for poorly-rated candidates, $p < 0.06$). All p-values are computed using 1-way ANOVA.

A contrasting effect is seen in the speaking style of the interviewer, as they interact with these highly-rated candidates. When speaking to the candidates with higher ratings, the same interviewers on average, had higher levels of conversational engagement (mean = 0.1 for highly-rated candidates, mean = 0.06 for poorly-rated candidates, $p < 0.01$). As mentioned previously, conversational engagement is calculated as the coupling between interacting Markov chains representing the turn-taking dynamics.

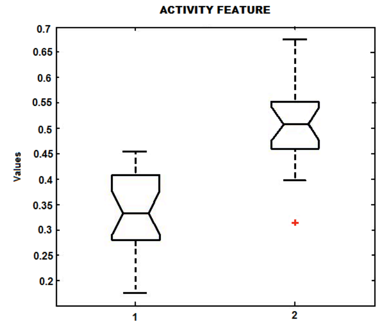

Figure 1. Candidates who perform better in interviews on average have higher levels of vocal activity ($p < 0.0001$) and lower levels of vocal emphasis ($p < 0.06$)

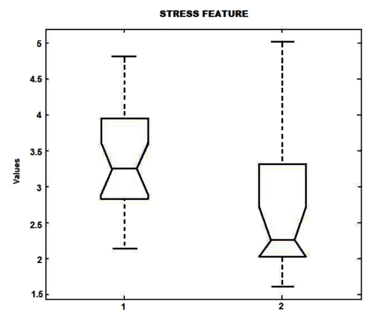

Figure 2. When speaking to highly-rated candidates, interviewers show higher engagement levels, as modeled using turn-taking dynamics ($p < 0.01$)

4.2 Predictive Model of Interview Outcomes

It is possible to use these differences in speaking-styles for better performing candidates to predict the performance of the candidate during the interview conversation. The *overall rating* given by the interviewer (scale of 1-5) has a strong correlation with the interviewee's speech features (R sqr = 0.67, $p < 0.0001$), and a slightly lower correlation with the interviewers speaking style (R sqr = 0.42, $p < 0.001$). Of the four sets of measures, the vocal activity and emphasis measures (R sqr = 0.71, $p < 0.00001$) are better predictors of *overall rating* than the engagement and mirroring features (R sqr = 0.41, $p < 0.001$).

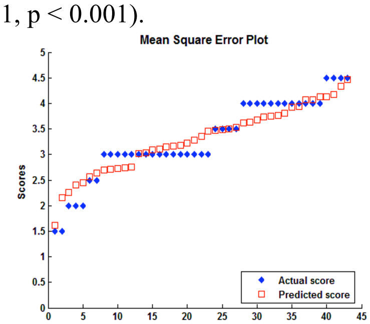

Figure 3. Mean Square Error (MSE) between predicted scores and actual scores for the overall rating category. The residual error is 0.1965

The ... the inte... aking styl... ion is

approximately the same the interviewee's extracted speech features (R sqr = 0.45, $p < 0.0001$) and the interviewer's extracted speech features (R sqr = 0.44, $p < 0.0001$). Similar to the previous case, the activity and emphasis features outperform the mirroring and engagement features.

As shown in the previous section, we can create a two-class model based on the *overall-rating* response (for highly-rated candidates, overall rating > 3). We use two classification approaches to predict whether a candidate would be 'highly-rated' or 'poorly-rated' using the extracted speech features. The model accuracy, precision, recall and f-measure results are in the table below. As would be expected, a Bayesian Network implementation out performs the Naïve-Bayes model in overall classification accuracy. Both models were validated using 4-fold cross-validation on the entire dataset.

Naïve Bayesian classifier 78% cross-validation accuracy overall			
	Precision	Recall	F-Measure
Highly-rated candiates	0.842	0.727	0.78
Poorly-rated candidates	0.739	0.85	0.791
Bayesian Network classifier 88% cross-validation accuracy overall			
	Precision	Recall	F-Measure
Highly-rated candiates	0.87	0.909	0.889
Poorly-rated candidates	0.895	0.85	0.872

5. DISCUSSION

In this paper, we demonstrate that the ongoing *social signaling* between an interviewee and interviewer during a job interview can be modeled using vocal tone and prosody features. The extracted features have a high correlation and predictive value with regard to the *overall rating* and *engaged in discussion* ratings. This implies that not only is there a non-linguistic communication channel in the job interview scenario, but that that it can be accurately quantified using automated methods, and even used as a basis to predict future outcomes.

Our results suggest that it may be possible to develop next-generation real-time software, off-line software and training methodologies that help interviewees to improve their communication skills, and provide better decision making tools to interviewers.

A key limitation of our work is the limited amount of labeled interview data available. We expect that our group and other researchers will continue to pursue this analysis with larger datasets and across different demographics.

Our work re-opens the mutual causality question for social psychologists. Is the speaking style simply reflective of the confidence of an interviewee, or does it have a causal relationship with a better outcome?

6. ACKNOWLEDGEMENTS

We would like to thank Shobhit Grover, Yogesh Wani and Anagha Gole for their help with the experiment design and data collection. We would also like to thank Dr. Alex (Sandy) Pentland for his insight and feedback.

7. REFERENCES

[1] R.D. Arvey, J. E. Campion, "The Employment interview: A summary and review of recent research", Personnel Psychology 35(2),281 - 322

[2] S. Basu, "Conversation Scene Analysis", in Dept. of Electrical Engineering and Computer science. Doctoral, MIT, 2002

[3] T. DeGroot, J. Gooty, "Can Nonverbal Cues be Used to Make Meaningful Personality Attributions in Employment Interviews?" Journal of Business and Psychology,2009

[4] R. Gifford, C. Fan Ng, & M. Wilkinson, "Nonverbal cues in the employment interview: Links between applicant qualities and interviewer judgments." , Journal of Applied Psychology, 70, 729-736, 1985

[5] A. Madan, R. Caneel. and A. Pentland , "Voices of Attraction", presented at Augmented Cognition, HCI, Las Vegas,2005

[6] A. Madan, R. Caneel, S. Pentland, "VibeFones: Socially aware Mobile Phones", International Symposium of Wearable Computing (ISWC), Switzerland, 2006.

[7] A. Pentland, "Social Dynamics: Signals and Behavior", ICDL, San Diego, CA, Oct 20-23,2004

[8] A. Pentland, J. Curhan, J. Khilnani, M. Martin, N. Eagle, R.Caneel, A. Madan, "A Negotiation Analyzer" appears UIST, Santa Fe, NM, Oct. 25-27 2004

[9] R. E. Riggio & B. Throckmorton, "The relative effects of verbal and nonverbal behavior, appearance, and social skills on evaluations made in hiring interviews.", Journal of Applied Social Psychology, 18(4), 331–348,1988