On Job Training: Automated Interpersonal Behavior Assessment & Real-Time Feedback

Skanda Muralidhar
Idiap Research Institute & EPFL, Switzerland
skanda.muralidhar@idiap.ch

ABSTRACT
In the service industry, customers often form first impressions about the organization based on the behavior, perceived personality, and other attributes of the front-line service employees they interact with. Literature has demonstrated that as a major component of interpersonal communication, nonverbal behavior (NVB) contributes towards shaping the outcome of these interactions and thus is key to determine customer satisfaction and perceived service quality. In this doctoral thesis, we aim to develop a computational framework for hospitality students with the goal of providing feedback for improving the impressions that other people make about them. Towards this, we collected a dataset of 169 laboratory sessions consisting of two role-plays, job interviews and reception desk scenarios, a total of 338 interactions. We present our approaches, results, works in progress, and planned future directions on these problems.

KEYWORDS
Social Computing; Job Interview; Nonverbal Behavior; First Impressions; Multimodal Interaction; Real-Time Feedback; Wearable Devices; Ubiquitous Computing; Google Glass

1 MOTIVATION
First impressions are crucial in all walks of life, be it a first date, making friends at a new school, a successful sales pitch, or even the acceptance of this paper. In psychology, first impression is defined as “the mental image one forms about something or someone after a first encounter”. Humans make initial judgments about a person’s attractiveness, likability, trustworthiness, competence and aggressiveness within one tenth of a second when meeting someone, and all these are based on nonverbal communication signals [1, 14, 31]. These initial judgments are especially important in formal settings like workplaces as it has been shown that these can influence crucial outcomes such as being hired or promoted. Good first impressions are particularly critical in sectors like sales, marketing and hospitality. Hence, this research investigates the interplay of nonverbal behavior and first impressions in two scenarios relevant for future careers of hospitality students, i.e. job interview and reception desk.

Existing literature in organizational psychology, nonverbal communication and hospitality indicates the existence of link between nonverbal behavior (NVB) and first impressions in workplaces. Traditionally, research in these fields has relied on manual coding of verbal and nonverbal cues due to the lack of tools and methods to automatically understand favorable impressions and related variables. Thus, making such studies labor-intensive and not easily scalable to either large number of users or different scenarios. With the explosion of unobtrusive sensory devices in the last decade, it has become increasingly easier to sense and interpret social constructs using user’s nonverbal cues computationally. The development of computational models has led to the incorporation of tools and technologies to automatically assess and analyze interpersonal behavior, known as Social Signal Processing [6, 22, 23, 29].

This new research area utilizes audio-visual processing and machine learning as additional analytical tools to automatically measure and analyze human behavior, with the aim of providing feedback (real-time or offline) to modify ones NVB in order to make favorable impressions. Existing psychology literature indicates that social interaction skills can be improved by practicing both verbal and nonverbal communication including how much, how fast, and how loud to talk, and how to regulate turn taking [8]. Advances in ubiquitous and wearable computing are enabling new possibilities to deliver real-time feedback [4, 7, 30].

The global aim of this doctoral research is to develop computational models to automatically analyze the relationship between NVB, behavioral first impressions, hirability and other related variables (Figure 1) and can be split into three broad objectives:

- **Objective-1**: Study the interplay of NVB and first impressions under two different settings (i.e. job interview & reception desk).
- **Objective-2**: Develop a real-time feedback system and evaluate its effect on dyadic interaction and impression formation.

Figure 1: Illustration of the idea of this proposal. The green dotted lines indicate the relationships under investigation in this thesis.
Table 1: Literature on behavioral analysis of human traits in various dyadic settings.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Social Variable</th>
<th>Best performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hung, 2007 [10];</td>
<td>Dominance</td>
<td>80.8%</td>
</tr>
<tr>
<td>Jayagopi, 2008 [12]</td>
<td></td>
<td>85%</td>
</tr>
<tr>
<td>Salamin, 2010 [25]</td>
<td>Role recognition</td>
<td>89%</td>
</tr>
<tr>
<td>Kim, 2012 [13]</td>
<td>Conflict levels in debates</td>
<td>R=0.80</td>
</tr>
<tr>
<td>Nguyen, 2014 [18]</td>
<td>Hirability, Big-5 personality</td>
<td>R^2=0.36 for hirability; up to 0.27 for other variables</td>
</tr>
<tr>
<td>Brilman, 2015 [2]</td>
<td>Successful Debate</td>
<td>75% for individual; 85% for team</td>
</tr>
<tr>
<td>Naim, 2015 [17]</td>
<td>Overall rating, Hiring decision &amp; others</td>
<td>R=0.70 for overall rating; up to 0.81 for other variables</td>
</tr>
<tr>
<td>Chen, 2016 [3]</td>
<td>Hiring decision; Big-5 traits</td>
<td>R=0.45 for hiring decision; up to 0.45 for other variables</td>
</tr>
<tr>
<td>Muralidhar, 2016 [16]</td>
<td>Overall impression, Big-5, Others</td>
<td>R^2=0.32 overall impression; up to 0.34 for other variables</td>
</tr>
<tr>
<td>Nguyen, 2016 [20]</td>
<td>Hirability, Big-5, Social, Communication and Professional skills</td>
<td>R^2=0.27 for Extraversion; up to 0.20 for social skills</td>
</tr>
</tbody>
</table>

and vocal modality like laughter, speech intonation etc. The multi-disciplinary domain of Social Signal Processing involving speech processing, computer vision, machine learning, and ubiquitous computing has been a field of growing interest within the multimedia community [6, 22, 23, 28, 29]. Various social situation have been investigated with an aim to predict understand diverse social constructs from automatically extracted nonverbal cues. For example, [10, 12] investigated dominance in group meeting using various audio-visual nonverbal cues extracted automatically.

3 NOVELTY
The overall contributions of this thesis are as follows
• We collect a novel dataset of 169 video in two different settings (job interview and hotel reception desk).
• We study human behavior and first impressions in a reception desk scenario, a setting not studied in computing literature.
• We compare and contrast two different setting to understand the role of context in human behavior.
• We report gender differences in impression formation which has implications for psychologists and human resources research.
• We develop a real-time behavioral awareness tool using Google Glass without negatively influencing dyadic social interaction.

4 APPROACH
In this PhD work, interpretability of machine-extracted behavioral cues has been given a high priority and is due to (a) close collaboration with social scientists, who are interested in explanations and processes rather just performance numbers (b) our interest to understand which nonverbal cues are important from the feedback point of view. To meet the objectives of this doctoral thesis, we collect two datasets and are briefly detailed in the following sections. The analysis of reception desk data collected as part of Objective-1 is currently work in progress.

4.1 Objective-1
Towards this objective, we collected a novel corpus consisting of 169 dyadic interaction in two settings (job interview and hotel front desk) [16]. This dataset, the UBImpressed dataset (Figure 2), was collected by embedding the experiments in the daily school routine of the hospitality students. The choice of the settings was motivated by the fact that these scenarios are of primary importance to hospitality students. The premise of the interview role-play was that the participant was applying for an internship at a luxury hotel. Structured interview was used, i.e., each interview followed the same sequence of questions so that comparisons across subjects could be made, as they are amongst the most valid tools for selecting applicants according to literature in psychology [9]. The setup was recorded using two Kinect v2 devices, one for each protagonist in the interaction. Audio was captured at 48kHz using an array of microphones that automatically performs speaker segmentation based on sound source localization. Cross-sensor synchronization was obtained by manually adjusting the delay between the modalities.

In the next step, various visual and auditory NVB cues (like prosody, speaking time, head nods) were automatically extracted from both the participant and interviewer. The choice of nonverbal cues were guided by their relevance in existing literature in social psychology [5, 11] and social computing [18, 19]. In parallel, a number of variables was manually labelled by five independent raters to act as ground truth. This list of variables includes hirability and first
impressions, a detailed list can be found in [15]. The raters watched the first two minutes of the videos and rated a number of social variables on a Likert scale from 1 (min) to 7 (max). The use of thin slices is a common practice in psychology [1] and social computing [24].

In the next step, a correlation analysis was conducted to understand the interplay between the various extracted features and the ground truth. We then defined inference of ground truth from the features extracted in a regression task using machine learning algorithms.

4.2 Objective-2

For this, we designed and implemented an automatic, real-time, conversational behavior awareness system for young sales apprentices making them aware of their NVB during customer interaction. As the human brain is not adept at multitasking, any significant distraction might lead to behavioral artifacts like stuttering, awkward pauses or smiles [21]. Thus, providing feedback during a dyadic interaction without negative impact is a challenge. Additionally, continuous staring at the feedback screen might lead to losing eye contact with the protagonist, which in turn causes a degradation in quality of interaction. Considering these constraints, we developed a pilot Google Glass (GG) app for real-time behavior awareness [15].

This GG based tool consists of two main components; sensing and feedback. The sensing component is responsible for perceiving and processing the user’s nonverbal behavior for which the built-in microphone of GG was exploited (Figure 3a). The feedback generation component uses the resulting analysis and presents the appropriate messages either visually or aurally (Figures 3b & 3c). The choice of nonverbal cue was motivated by results from literature in psychology [14] and other hardware constraints [15].

To evaluate the design and usefulness of the app, we conducted a pilot study with 15 sales apprentices from a local vocational education and training (VET) school [15]. The interaction consists of a typical sales scenario in a mobile phone shop (average duration = 2.5 minutes). In this scenario, the participant played the role of salesperson and wore the GG. During the interaction, GG would provide automatic feedback on behavioral cues and to follow the suggestion or not was left to the discretion of the participant. The role of the client was played by a researcher who was a native French speaker with directions to elicit two specific behaviors from the participants (a) talk for a relatively long time (b) remain silent.

To assess the impact of glass on dyadic interaction, the interaction videos were annotated by two groups of native French speakers.

Group-A and Group-B consisted of two and three raters respectively. Group-A was informed, at length, about GG and the feedback provided by it, while Group-B was not. For both groups, the part of the screen which displayed feedback was blocked. Group-A was asked to watch the video and answer: Do you believe the salesperson was given feedback, based on the behavior of the person throughout the video? in the form of Yes, No or Maybe. Annotators in Group-B were asked to rate the video on a five-point Likert scale (1 = ‘very poor’ to 5 = ‘very good’). Specifically they were asked Consider yourself to be the client in this interaction and rate the participant on (a) Overall performance (OP) (b) Quality of interaction (QoI).

5 EXPERIMENTS & RESULTS

5.1 Objective-1

The preliminary correlation analysis indicated a positive correlation between speaking time and overall impression, while silence events was negatively correlated. This indicated that participants who spoke more were rated higher. In the next step, a regression task to predict the overall impression with nonverbal cues as predictors was defined and the results are summarized in Figure 4. For evaluation of performance, we utilized the coefficient of determination $R^2$, a metric often used in both psychology and social computing. $R^2$ accounts for the amount of total variance explained by the regression model under analysis. Results from utilizing all the data points indicate that overall impression annotated was predictable to some degree from nonverbal behavior ($R^2=0.32$). This implies nonverbal behavior is predictive of overall first impression as shown in existing literature [14] and corroborates the results found in [18].

Comparing this results to recent works, in [18], the authors reported $R^2=0.36$ for hirability, a measure we have not used. Naim et al. reported results on a different set of social variables using correlation coefficient ($r$) as their evaluation metric [17]. We compare our results to this work by converting $r$ to $R^2$ (our evaluation metric, coefficient of determination $R^2$ is obtained by computing the square of correlation coefficient $r$). They reported a prediction accuracy of $r=0.70$ for overall performance, which indicates a $R^2=0.49$. There is no direct way of assessing where the performance difference come from, as the dataset used is not publicly available to our knowledge.
Although no significant difference could be observed between males and females in terms of the values of annotations, we observe that interviews with male participants were predicted with higher accuracy ($R^2=0.44$) than the ones featuring females ($R^2=0.06$). To understand these differences, we analyzed the correlations between nonverbal cues and the annotated variables for data subsets separated based on gender, the results are discussed in detail in [16].

5.2 Objective-2
The distribution of annotation data for both variables is presented in Figure 5. Median rating of OP is 3.25 (max= 5; min= 1), while the median rating of QoI is 3 (max= 4.5; min= 1). Due to the limitations in dataset size, no firm statistical conclusions can be drawn for social variables. Also, talking more does not imply a better conversation. Another limitation of this work is that the QoI and OP was not validated by domain experts (speaking coach or sales coach).

Figure 6 indicates that for majority of the videos, the annotators of Group-A were unable to correctly infer if feedback had been provided (adding the “no” and “maybe” columns in Figure 6). These results suggest that in several cases the reaction of GG users to the feedback is either subtle or does not deviate from what an external observer would consider as usual conversational behavior.

To investigate this issue in more detail, the behavior of participants during the interaction was manually coded by the authors to understand how subjects react to real-time feedback. The manual coding of behavior signal that some subjects smiled or giggle when feedback was provided, possibly due to both the actual experience of receiving feedback combined with a novelty effect. Reactions to both types of feedback and time to heed to suggestion is presented in Table 2. This in conjunction with annotations by Group-A on inference of feedback (Figure 6) indicate that in the majority of the cases reaction to feedback was natural.

Table 2: Behavioral reactions to feedback. Time to heed is the time to taken to accept the feedback i.e stop talking if feedback says stop talking.

<table>
<thead>
<tr>
<th>Feedback Type</th>
<th>Reaction</th>
<th>Time to heed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speak Less</td>
<td>Smiling, Laughing, Squinting</td>
<td>1-4 seconds</td>
</tr>
<tr>
<td>Speak More</td>
<td>Smiling</td>
<td>2-4 seconds</td>
</tr>
</tbody>
</table>

6 WORK IN PROGRESS
We have investigated the role of nonverbal behaviour in the formation of first impressions in the much studied dyadic setting of employment interviews. Current on-going work, investigates another important aspect of this interaction; verbal communication. It have been shown in literature that use of words in correlated to emotional state being experienced by the person [27]. Authors in [17] have shown that students who used “we” instead of “i” were found to be more hirable. Thus, we want to understand linguistic style and its implications for impressions in hospitality industry.

We are also currently exploring an interesting and novel setting not studied previous in computing literature: hotel reception desk. This is an important form of interaction in the hospitality industry as the reception desk is the “face” of the hotel and customers make their first impressions of the establishment based on their interactions with reception desk assistants [26]. Hence, good interpersonal communication of the reception desk assistants is of great importance. We also plan to understand impression formation in this novel setting by extracting various nonverbal and verbal behavioural features and investigating their correlation to impression formation.

7 FUTURE WORK
So far, we have explored the use of traditional hand-crafted features and machine learning algorithms to build a computational framework to infer first impressions, hirability and related traits. In the next part of my doctoral research, we would like to explore the use of deep learning methods - the state of the art for features representation in many machine recognition tasks - to automatically infer hirability and related trait impressions in job interviews. This motivated by the fact that the emergence of deep learning gives an opportunity to improve performance in automated tasks.

ACKNOWLEDGMENTS
This work was funded by the UBIMPRESSED project of the Sinergia interdisciplinary program of the Swiss National Science Foundation (SNSF). The author would like to thank Daniel Gatica-Perez, Mari-anne Schmid Mast, Laurent Son Nguyen for discussions and advice, Denise Frauendorfer and the research assistants for their help in data collection.
REFERENCES


