Social Sensing Methods for Analysis of Dyadic Hospitality Encounters

Thèse N° 9177

Présentée le 4 avril 2019

à la Faculté des sciences et techniques de l'ingénieur
Laboratoire de l'IDIAP
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

Skanda MURALIDHAR

Acceptée sur proposition du jury

Dr D. Gillet, président du jury
Prof. D. Gatica-Perez, directeur de thèse
Prof. M. Schmid Mast, rapporteuse
Prof. C. PELACHAUD, rapporteuse
Dr P. Pu, rapporteuse

2019

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

To my grandmothers and parents...

Acknowledgements

"*We are like dwarfs on the shoulders of giants, so that we can see more than they, and things at a greater distance, not by virtue of any sharpness on sight on our part, or any physical distinction, but because we are carried high and raised up by their giant size.*"

–Bernard of Chartres, 1130 AD

I have the good fortune of being held up on the shoulders of many giants in my life, and at this important milestone in my life, I would like to thank them all.

Firstly I would like to thank Prof Daniel Gatica-Perez for all his help, patience, and guidance as I fumbled from one mistake into another during the last four years. Without your advice, this thesis would not have been possible. There is a saying in Sanskrit that a student learns more from the nonverbal behavior of ones teacher than his words. After four years working with you, I have a deeper understanding of this quote. Apart from professional advice, I would also like to thank you for advice in my personal life which has led me to be a better person. I would also like to thank other members of my committee Dr. Denise Gillert, Dr. Pearl Pu, Prof. Marianne Schmid Mast and Prof. Catherine Pelachaud for taking time to review this document and for their comments and feedback. I would like to acknowledge that this thesis was funded by the UBIMPRESSED project of the Sinergia interdisciplinary program of the Swiss National Science Foundation (SNSF).

I would like to thank all my group members current and former. Laurent, thank you for patience and advice when I was struggling in the first year. You have been a source of strength and knowledge I could bank on when I felt lost during the initial years. Darshan bhai, thank you for your discussions and being there to bounce ideas off. You and Neha made me feel like family in Martigny. Joan, thank you for your advice, great critic and discussions. The advice and discussions during lunch at DH lab were valuable. I would like to the other members of the group, Rui, Gulcan, Trung, Oya, and Dayra for helping me out at various times and for your support when the going got difficult. I would also like to thank Denise, Ailbhe and all the RA who helped in the data collection, Jean-Marc, YuYu, Kenneth and Remy for all your help in the duration of the project.

I would like to thank others in 305 office who made my stay there super fun. Thank you, Pranay, Dhananjay, Shubadeep, Vinayak and Bastien for all the technical ideas, time-pass discussions

over coffee and fun times outside the office including hikes, SAMs and cricket. I would also like to thank Leslie, Suraj, Banri and Pavan for the interesting conversations, varied points of view of my research. I would like to Mathew for all this advice, suggestions and lunch discussions. These were of great help in some of the research ideas implemented in this thesis. I would also like to thank Frank, Bastien, Louis-Marie, and others in the 405 for always being there to help out in all my technical issues. I would like to thank Nadine and Sylvie for their help in the administrative work, paperwork for conference travel and to liaise with the commune in French.

Outside of Idiap, I would like to thank all my friends at EPFL, Jessica, Marie, Wissam it was fun attending classes and doing course projects with you guys. I would like to thank my other friends in Lausanne, Alex, Lucie, David, Clara, and Jesus. You guys were great to hang out with and to share an apartment with. I will miss the winter fondue parties and playing fetch with Lenni. Thank you, Vijay and Priti for being so welcoming and inviting me for home cooked Indian food because of which I never suffered any homesickness. Vijay, your pep talks pushed me into setting high goals and set out to achieve them.

I would like to thank my other teachers, Prof. Mohan Kankanhalli (NUS School of Computing) and Prof. K. R. Ramakrishan (Department of Electrical Engineering, Indian Institute of Science) for their implicit encouragement and faith in my abilities, always available to discuss any problem scientific or otherwise. Another person I remain indebted to is Mr. Bhaskar, without whose constant faith in my abilities and encouraged to push myself beyond my comfort boundaries. Our discussions about the material world and Indian philosophy have made me a well-rounded individual.

I would like to thank my closest friends Sreetama and Sarvagna being there for me when I needed support or when I was feeling down. I would especially like to thank Sreetama for her help in proofreading all my papers, research discussions and helping me set high ambitions in life. At this time, I would like to remember and thank my grandmothers, who took care of me, instilled values and encouraged me to seek out my own path in this world. Last but not least I would like to thank my parents, my brother and my family for their unconditional love and support. Thank you, mother and father, for all the sacrifices you have made to give me a better chance in life.

Lausanne, 29 October 2018 Skanda Muralidhar

Abstract

First impressions are critical to professional interactions, especially in service industry like hospitality. In the service industry, customers often assess quality of service based on the behavior, perceived personality, and other attributes of the front-line service employees they interact with. Interpersonal communication during these interactions is thus key to determine customer satisfaction and perceived service quality.

Given the importance of first impressions in hospitality, this thesis contributes to the implementation of a behavioral training framework for hospitality students with an aim of improving the impressions that other people make about them in workplaces. We outline the challenges associated with designing such a framework and embedding it in the everyday practice of a real hospitality school. These behavioral training sessions were recorded based on principles of unobtrusive measurements and social signal processing. We collect a dataset of 169 laboratory sessions consisting of two role-plays; job interviews and reception desk scenarios, for a total of 338 interactions.

In our first study of the job interview scenario, we evaluate the relationship between automatically extracted verbal and nonverbal cues with the various manually annotated impressions of social variables in a correlation analysis. We then develop methods to automatically infer first impressions using verbal cues, nonverbal features and their combination. Our inference results indicate that nonverbal features outperform verbal cues in an inference task. Best inference performance is obtained by fusion of verbal and nonverbal cues. A gender based analysis reveals important differences between males and females in terms of nonverbal cues displayed and impressions formed. This result is corroborating previous findings in psychology.

In our second study we investigate the reception desk interaction. We aimed to develop a computational framework to automatically infer perceived performance and skill variables using nonverbal and verbal behavior displayed. We also study the connections between receptionists' impressions of Big-5 personality traits, attractiveness, and performance. Our results indicate the feasibility of inferring perceived job performance from nonverbal and verbal cues displayed. Furthermore, contrary to our hypothesis, perceived attractiveness had low predictive power of first impressions.

We then conduct a cross-situation analysis to understand the impact of situation in the formation of first impressions. This is based on the truism in psychology that same people behave differently in different situations. A correlation analysis reveals connection between perceived variables and nonverbal cues displayed during job interviews, and perceived performance on the job. We develop a computational framework to infer the perceived performance and soft skills in the reception desk situation with nonverbal cues and linguistic style from the two interactions as predictors. The best inference performance is achieved by fusing nonverbal cues displayed during the reception desk setting and the human-rated interview scores. We observe that some behavioral cues (greater speaking turn duration and head nods) are positively correlated to higher ratings for all perceived variables across both situations. This is one of the major contributions of this thesis.

Our analysis of the two situations individually and cross-situation has shown the importance of speaking time in the formation of first impressions in dyadic interactions. Using this knowledge, we designed and developed a Google Glass based system to provide unobtrusive, real-time behavioral awareness to the user.The effectiveness of this system is then evaluated in a pilot study consisting of 15 apprentices from a vocational education and training (VET) school. The users found the system to be fun to use, little distracting and useful. A manual coding of the recorded videos indicated that the dyadic interaction was not negatively influenced by the real-time feedback.

Verbal channel of interpersonal communication is a seldom investigated behavioral cue. This could be attributed to manual transcriptions of social interactions being a time consuming and expensive process. Applying deep learning based advances in speech recognition and natural language processing, we investigate the role of verbal behavior in how we are perceived by others. Towards this aim, we use noisy real-world videos resumes from YouTube. Video resumes are short videos in which job applicants present themselves and their communication skills to potential employers. Such videos have grained popularity with the wide-spread popularity of social media like YouTube. We aim to (a) identify the best representation for verbal content to infer first impressions, (b) measure the impact of automatic speech recognition on inference performance as compared to manual transcription. Our results indicate the feasibility of using verbal content in inferring first impressions in online conversation video resumes using manual transcription. We also observe that the performance of ASR is as good as those obtained using manual transcription. Our work indicate the feasibility of using ASR for transcribing social interactions to study work-related social constructs and first impressions at large scale.

Key words: social computing, first impressions, hospitality, nonverbal behavior, verbal content, multimodal interaction, hirability, job performance, reception desk, online video resume, Doc2Vec, Word2Vec, GloVe, real-time feedback, wearable devices, ubiquitous computing, google glass

Zusammenfassung

Der erste Eindruck ist entscheidend für die professionelle Interaktion, insbesondere in der Dienstleistungsbranche wie der Gastronomie. In der Dienstleistungsbranche bewerten Kunden die Servicequalität oft anhand des Verhaltens, der wahrgenommenen Persönlichkeit und anderer Attribute der Mitarbeiter mit denen sie an vorderster Front zusammenarbeiten. Die zwischenmenschliche Kommunikation während dieser Interaktionen ist daher der Schlüssel zur Ermittlung der Kundenzufriedenheit und der wahrgenommenen Servicequalität.

Angesichts der Bedeutung des ersten Eindrucks in der Gastronomie trägt diese These zur Umsetzung eines Verhaltens-Training-Rahmens für Studenten der Gastronomie mit dem Ziel der Verbesserung der Eindrücke, die andere Menschen über sie in der Arbeitswelt haben, bei. Wir skizzieren die Herausforderungen, die mit der Gestaltung eines solchen Rahmens und seiner Einbettung in die tägliche Praxis einer echten Gastronomie Schule verbunden sind. Diese Verhaltenstrainingseinheiten wurden nach dem Prinzip der unauffälligen Messungen und der sozialen Signalverarbeitung aufgezeichnet. Wir sammeln einen Datensatz von 169 Laborsitzungen, bestehend aus jeweils zwei Rollenspielen: Vorstellungsgesprächen und Szenarien am Empfang. Insgesamt 338 Interaktionen.

In unserer ersten Studie des Vorstellungsgesprächsszenarios bewerten wir den Zusammenhang zwischen automatisch extrahierten verbalen und nonverbalen Hinweisen mit verschiedenen manuell kommentierten Indikatoren von sozialen Variablen in einer Korrelationsanalyse. Anschließend entwickeln wir Methoden um aus verbalen Hinweisen, nonverbalen Merkmalen und deren Kombination automatisch den ersten Eindruck abzuleiten. Unsere Ergebnisse deuten darauf hin, dass nonverbale Merkmale verbale Merkmalen in einer Inferenz-Aufgabe übertreffen. Die beste Inferenzleistung wird durch die Kombination von verbalen und nonverbalen Hinweisen erzielt. Eine geschlechtsspezifische Analyse zeigt wichtige Unterschiede zwischen Männern und Frauen in Bezug auf nonverbale Hinweise und Eindrücke. Dieses Ergebnis bestätigt frühere Ergebnisse in der Psychologie.

In unserer zweiten Studie untersuchen wir die Interaktion an der Rezeption. Wir wollen eine Formel entwickeln um automatisch auf wahrgenommene Leistungs- und Qualifikationsvariablen unter Verwendung von nonverbalem und verbalem Verhalten zu schließen. Wir untersuchen auch die Zusammenhänge zwischen den Eindrücken des Rezeptionisten/der Rezeptionistin anhand Big-5-Persönlichkeitsmerkmalen, Attraktivität und Leistung. Unsere Ergebnisse zeigen die Machbarkeit der Ableitung der wahrgenommenen Arbeitsleistung aus den dargestellten nonverbalen und verbalen Hinweisen. Außerdem haben die wahrgenommene Attraktivität entgegen unserer Hypothese eine geringe Vorhersagekraft für den ersten

Eindruck.
Anschließend führen wir eine situationsübergreifende Analyse durch, um die Auswirkungen der Situation bei der Bildung des ersten Eindrucks zu verstehen. Dies basiert auf der Binsenweisheit in der Psychologie, dass sich dieselben Menschen in verschiedenen Situationen unterschiedlich verhalten. Eine Korrelationsanalyse zeigt den Zusammenhang zwischen wahrgenommenen Variablen und nonverbalen Hinweisen, die während der Vorstellungsgespräche auftreten, und der wahrgenommenen Leistung am Arbeitsplatz. Wir entwickeln eine Formel um die wahrgenommene Leistung und Soft-Skills in der Rezeptionssituation mit nonverbalen Hinweisen und Sprachstilen aus den beiden Interaktionen als Prädiktoren abzuleiten. Die beste Inferenzleistung wird durch die Kombination von nonverbalen Hinweisen, die während der Empfangssituation auftreten, und den von Menschen bewerteten Interview-Werten erreicht. Wir beobachten, dass einige Verhaltensmuster (längere Redezeit und Kopfnicken) positiv mit höheren Bewertungen für alle wahrgenommenen Variablen in beiden Situationen korreliert sind. Dies ist einer der wichtigsten Beiträge dieser Arbeit.
Unsere Analyse der beiden Situationen, individuell und situations-übergreifend, hat gezeigt wie wichtig die Redezeit bei der Bildung des ersten Eindrucks in dyadischen Interaktionen ist. Mit diesem Wissen haben wir ein auf Google Glass basierendes System entwickelt, um dem Benutzer ein unauffälliges Echtzeit-Verhaltensbewusstsein zu vermitteln, dessen Effektivität in einer Pilotstudie mit Auszubildenden aus einer Berufsbildungsschule für je 15 US-Dollar bewertet wird. Die Benutzer fanden das System lustig in der Anwendung, wenig ablenkend und nützlich. Eine manuelle Codierung der aufgezeichneten Videos zeigte, dass die dyadische Interaktion nicht negativ durch die Echtzeitrückmeldung beeinflusst wurde.
Der verbale Kanal der zwischenmenschlichen Kommunikation ist ein wenig erforschter Verhaltenshinweis. Dies könnte darauf zurückzuführen sein, dass die manuelle Transkription von sozialen Interaktionen ein zeitaufwändiger und kostspieliger Prozess ist. Durch die Anwendung Deep-Learning-basierter Spracherkennung und natürlichen Sprachverarbeitung untersuchen wir die Rolle des verbalen Verhaltens bei der Wahrnehmung durch andere. Zu diesem Zweck verwenden wir verrauschte, realistische Videos, die von YouTube stammen. Video-Lebensläufe sind kurze Videos in denen sich Bewerber und ihre Kommunikationsfähigkeiten potenziellen Arbeitgebern vorstellen. Solche Videos haben hohe Popularität mit der weiten Verbreitung von Social Media wie YouTube gewonnen. Unser Ziel ist es, (a) die beste Darstellung für verbale Inhalte zu ermitteln, um erste Eindrücke zu erhalten, (b) die Auswirkungen der automatischen Spracherkennung (ASR) auf die Inferenzleistung im Vergleich zur manuellen Transkription zu messen. Unsere Ergebnisse zeigen die Machbarkeit der Verwendung verbaler Inhalte bei der Ableitung erster Eindrücke in Online-Konversationsvideos auf Grund manueller Transkription. Wir stellen auch fest, dass die Leistung von ASR genauso gut ist wie die, die durch manuelle Transkription erzielt wird. Unsere Arbeit zeigt die Machbarkeit der Verwendung von ASR für die Transkription sozialer Interaktionen zur Untersuchung arbeitsbezogener sozialer Konstrukte und erster Eindrücke im großen Maßstab.

Stichwörter:

Résumé

La première impression est critique lors des interactions professionnelles, et plus particulièrement dans les professions de service comme l'hôtellerie. Dans ce domaine, les clients évaluent la qualité du service en se basant sur le comportement, la personnalité perçue et d'autres attributs du personnel avec lequel ils entrent en contact. Durant ces interactions, la communication interpersonnelle est ainsi un facteur clé de la satisfaction des clients et de la qualité du service perçue.

Etant donné l'importance de la première impression dans le secteur hôtelier, cette thèse contribue à l'élaboration d'un cadre de formation pour les étudiants en hôtellerie pour perfectionner leur attitude et leur comportement dans le but d'améliorer l'impression qu'ils font aux clients sur leur place de travail. Nous avons mis en évidence les défis associés à la création d'une telle formation et l'avons intégrée dans la pratique quotidienne d'une école hôtelière. Ces sessions de formation comportementale furent enregistrée en suivant les principes de l'enregistrement non-intrusif et du traitement de signaux sociaux. Nous avons collecté une base de données de 169 sessions en laboratoire constituées de deux scénarios - un entretien d'embauche et un bureau de réception - pour un total de 338 interactions.

Dans notre première étude du scénario d'entretien d'embauche, nous avons extrait automatiquement les composantes verbales et non-verbales et nous avons évalué leur corrélation avec l'impression qu'elles donnent par rapport à différentes variables sociales, annotées manuellement. Nous avons ensuite développé des méthodes qui déduisent automatiquement ces premières impressions en se basant sur les composantes verbales, non-verbales ainsi que leur combinaison. Les performances obtenues avec ces méthodes indiquent que les caractéristiques non-verbales sont plus importantes que les caractéristiques verbales pour ce genre de tâche. Cependant, les meilleurs résultats furent obtenus en combinant les deux. Aussi, notre analyse a montré des différences importantes entre les hommes et les femmes en termes de signaux non-verbaux et de l'impression donnée. Ces résultats vont dans le même sens que plusieurs études psychologiques précédentes.

Dans notre deuxième étude, nous nous sommes concentrés sur le scénario du bureau de réception. Notre but était de développer une méthode pour déduire automatiquement des variables de performance et de niveau de compétence perçues en analysant le comportement verbal et non-verbal. Nous avons également étudié les liens entre l'impression donnée par le réceptionniste sur les types psychologiques Big-5, l'attirance exercée sur l'observateur et la performance. Nos résultats montrent la possibilité de déduire la performance perçue en utilisant les signaux verbaux et non-verbaux du réceptionniste. De plus, contrairement à notre

hypothèse, l'attirance montre une faible puissance prédictive pour la première impression.
Nous avons ensuite mené une analyse trans-scénario pour comprendre l'impact de la situation sur la formation de la première impression, basé sur l'évidence psychologique qu'une même personne se comporte différemment dans différentes situations. Une analyse de corrélation a révélé un lien entre les signaux non-verbaux et les différentes variables perçues durant les entretiens d'embauche et la performance professionnelle perçue. Nous avons développé une méthode pour déduire la performance perçue et les soft skills dans le cas du bureau de réception en utilisant comme prédicteurs les signaux non-verbaux et le style linguistique des deux scénarios. Les meilleurs résultats furent atteints en combinant les signaux non-verbaux dans le cas du bureau de réception et la performance évaluée par un observateur pendant l'entretien d'embauche. Nous avons observé que certains signaux (tours de paroles plus longs, hochements de têtes) sont positivement corrélés à l'évaluation de toutes les variables perçues dans les deux scénarios. Il s'agit de l'une des contributions majeures de cette thèse.
Notre analyse des deux scénarios pris individuellement et ensemble a montré l'importance du temps de parole dans la formation de la première impression lors des interactions dyadiques. Sachant cela, nous avons conçu et développé un système basé sur les Google Glass pour fournir à l'utilisateur un feedback non-intrusif sur son comportement en temps réel. Nous avons évalué l'efficacité de ce système dans une étude pilote avec 15 apprentis en formation professionnelle. Les utilisateurs ont trouvé le système amusant, peu distractif et utile. Un encodage manuel des vidéos enregistrées a indiqué que l'interaction dyadique n'était pas influencée négativement par ce feedback en temps réel.
Le canal verbal est un signal comportemental rarement étudié dans les communications entre individus, principalement à cause du temps important et du coût élevé que demande la transcription manuelle d'interactions sociales. En appliquant les avancées techniques du traitement de langage naturel basées sur le deep learning, nous avons étudié le rôle du comportement verbal dans notre perception des autres. Dans ce but, nous avons utilisé de véritables CV vidéos pris sur YouTube. Les CV vidéos sont de courtes vidéos dans lesquels une personne qui cherche un emploi se présente elle et ses compétences de communications aux potentiels employeurs. Ce genre de vidéo a gagné en popularité avec le développement des réseaux sociaux comme YouTube. Notre but était (a) d'identifier les meilleures représentations de contenu verbal pour déduire la première impression et (b) de mesurer l'impact de l'utilisation de la reconnaissance vocale automatique (RVA) plutôt que la transcription manuelle sur les performances de déduction de la première impression. Nos résultats indiquent qu'il est possible d'utiliser le contenu verbal d'une conversation en ligne ou d'un CV vidéo en se basant sur la transcription manuelle. Nous avons également observé que les performances des méthodes qui utilisent la reconnaissance vocale automatique sont comparables à celles qui emploient la transcription manuelle. Notre analyse indique donc la possibilité de l'utilisation de la reconnaissance vocale automatique pour transcrire les interactions sociales dans le but d'étudier la formation de la première impression et les constructions sociales liées au monde professionnel à grande échelle.

Mots clefs :

Contents

Acknowledgements **i**

Abstract (English/Français/Deutsch) **iii**

List of figures **xiii**

List of tables **xv**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Objectives . . . 3
1.3 Contributions . . . 4
1.4 Outcomes . . . 5
1.5 Organization of this Thesis . . . 6
1.6 Publications . . . 6

2 Related Work **9**
2.1 Literature in Psychology & Hospitality . . . 9
2.1.1 Job Interviews . . . 9
2.1.2 Perceived Performance . . . 10
2.1.3 Cross Situation Analysis in Literature . . . 11
2.2 Literature in Computing . . . 12
2.2.1 Job Interviews . . . 12
2.2.2 Perceived Performance . . . 13

3 Data Collection & Feature Extraction **15**
3.1 Context and Challenges . . . 15
3.2 Overall Design . . . 16
3.2.1 Lab Sessions . . . 18
3.2.2 Feedback Session . . . 19
3.2.3 Participants . . . 19
3.3 Annotations . . . 20
3.3.1 Job Interviews . . . 20
3.3.2 Reception Desk . . . 20
3.4 Transcriptions . . . 21

3.5 Extraction of Behavioral Cues . . . 22
3.5.1 Audio-Visual Nonverbal Cues . . . 22
3.5.2 Verbal Cues . . . 24
3.6 Conclusion . . . 25

4 First Impressions in Employment Interviews 27
4.1 Dataset . . . 28
4.1.1 Data Corpus . . . 28
4.1.2 Transcripts . . . 30
4.2 Verbal and Nonverbal Features . . . 30
4.3 Principal Component Analysis of Skills . . . 30
4.4 Correlation Analysis . . . 31
4.4.1 Overall Impression & Nonverbal Cues . . . 31
4.4.2 Overall Impression & Verbal Behavior . . . 34
4.5 Regression Analysis . . . 34
4.5.1 Inferring First Impressions from Nonverbal Cues . . . 35
4.5.2 Changes Across Sessions . . . 37
4.5.3 Inferring First Impressions from Verbal Cues . . . 38
4.5.4 Inferring Principal Components from Features . . . 39
4.6 Conclusion . . . 40

5 First Impressions in Reception Desk 43
5.1 Dataset . . . 44
5.1.1 Data collection & Annotations . . . 45
5.2 Verbal and Nonverbal Features . . . 45
5.3 Correlation Analysis . . . 45
5.3.1 Performance Impressions & Personality Trait Impressions . . . 47
5.3.2 Performance Impressions & Nonverbal Cues . . . 47
5.3.3 Performance Impressions & Attractiveness . . . 49
5.3.4 Performance Impressions & Verbal Cues . . . 50
5.4 Regression Analysis . . . 50
5.4.1 Inferring Performance Impressions from Nonverbal Cues . . . 52
5.4.2 Inferring Performance Impressions from Big-5 Impressions . . . 53
5.4.3 Inferring Performance Impressions from Attractiveness . . . 54
5.4.4 Inferring Performance Impressions from Verbal Cues . . . 55
5.5 Conclusion . . . 55

6 Cross-situation analysis 57
6.1 Data Corpus . . . 59
6.1.1 Annotations . . . 59
6.1.2 Speech Transcripts . . . 60
6.2 Nonverbal and Verbal Feature Extraction . . . 60
6.3 Inference Framework and Experimental Protocol . . . 61

6.4 Results and Discussion . . . 62
6.4.1 RQ1: Perceived Variables in Interview and Reception Desk Situations . . 63
6.4.2 RQ2: NVB in Interviews and Perceived Performance at the Desk . . . 66
6.4.3 RQ3: Linguistic Content and Perceived Performance . . . 69
6.4.4 Qualitative Study . . . 71
6.5 Conclusion . . . 73

7 Other Applications **75**
7.1 Verbal Content and Hirability Impressions in Video Resumes . . . 75
7.1.1 Related Work . . . 77
7.1.2 Dataset . . . 79
7.1.3 Method . . . 82
7.1.4 Results and Discussion . . . 84
7.1.5 Conclusion . . . 89
7.2 Dites-Moi: Wearable Feedback on Conversational Behavior . . . 90
7.2.1 Approach . . . 91
7.2.2 Evaluation . . . 94
7.2.3 Conclusion . . . 97

8 Conclusions **99**
8.1 Implications . . . 101
8.2 Limitations & Future Work . . . 102

Bibliography **117**

Curriculum Vitae **119**

List of Figures

3.1 Illustration of the design of the behavioral training procedure. 17

3.2 (a) Snapshot of the job interview situation collected as part of the behavioral training program; (b) Questions asked during job interview where the participant is role playing an applicant for an internship in a high-end hotel. 17

3.3 (a) Snapshot of the hotel reception desk situation collected as part of the behavioral training program; (b) Details of the desk situation where participants have to handle an unhappy client. 18

3.4 Gaze bias correction method using speaking turns. 24

3.5 Flow chart showing the steps followed for extracting Doc2Vec features from transcribed text of both the situations. 25

4.1 Clustering of perceived skills after principal component analysis (PCA). The first three principal components, accounting for 94.3% of the variance, are displayed here by projecting the original axes onto the PCA space ($N = 169$). 31

4.2 Difference in Overall Impression score between two lab sessions for all participants. 37

5.1 Wordcloud showing the frequency of words used during the 169 desk interactions. In this figure, the relative size of each word indicates its frequency. For example, in the English interactions, the most common word is *can*, while for French interactions it is *ca*. All words of the same color have the same size/frequency. 51

6.1 A visual summary of the cues used in our experiments and how they were obtained. 62

6.2 Box-plot showing the distribution of annotated scores for each of the variable of interest. Here we observe that mean scores for interview (yellow) is greater than mean scores for reception desk (blue). 63

6.3 List of questions sent to hospitality students as a part of a qualitative study to understand the implication of this work. 71

6.4 List of questions sent to hospitality school directos as a part of a qualitative study to understand the implication of this work. 72

7.1 Box plot illustrating the distribution of number of words obtained by (a) manual transcription [Man] (b) ASR for first 2 minutes [ASR-2min] (c) ASR for full video [ASR-Full] for a random subset of 292 videos. The dotted line indicates the mean value. 80
7.2 Overview of the work flow used in this study. The two classes of verbal content representation methods (a) document-based (b) word-based investigated is illustrated. For the document-based method, performance of LIWC and Doc2Vec in inferring hirability impressions, individually and combination is investigated. For the word-based method, all combinations of algorithm and aggregation techniques are investigated. 81
7.3 Overview of Google Glass sensors and visual feedback on Google Glass Display 91
7.4 View of the study setting: the participant works behind the desk; the interaction partner is not visible in the figure. 92
7.5 Distribution of participants' self ratings for (a) feedback modalities (higher is better for *useful* and *Ov. Impression*, lower is better for *Distract*) 95
7.6 Distribution of participants' self ratings for overall experience (higher is better) 96
7.7 Distribution of participants' self ratings for *overall performance* and *quality of interaction* ratings by annotators (higher is better) 96
7.8 Distribution of answers for prediction by Group-A. All Google Glass users received feedback. 98

List of Tables

4.1 Intra class correlation ($ICC(2,k)$) and descriptive statistics for perceived social variables in the job interview data corpus. The agreement between raters for all the social variables was greater than 0.5 indicating moderate reliability. Furthemore, the mean ratings of all social variables is greater then 4 indicting that the participants were overall percieved positively. 29

4.2 Correlation matrix for perceived social variables in job interviews (N=169) . In all cases, correlation is significant (p<0.01). 29

4.3 Selected Pearson's correlation coefficient for various social variables.* $p < 0.01$, † $p < 0.05$ 32

4.4 Range of Pearsons correlation between eye gaze, facial expressions and social variables in the interview ($N = 161$) dataset. *** $p < 0.001$;** $p < 0.01$;* $p < 0.05$. 33

4.5 Correlation between linguistic cues and impressions in job interviews ($N = 169$) with significance values ** $p < 0.001$; * $p < 0.05$. Others not significant. 34

4.6 Regression results for each language and gender using regression methods; pVal with random forest (pVal-RF) and pVal with Ridge (pVal-Ridge) * $p < 0.01$, † $p < 0.05$. 35

4.7 Correlation coefficients between selected nonverbal cues and overall impression for sub-sets separated based on gender. * $p < 0.01$; † $p < 0.05$ 36

4.8 Regression analysis using eye gaze and other visual features for job interviews ($N = 161$). 36

4.9 Descriptive stats of social variables and nonverbal cues. Mean values for speaking cues are after z-score normalization. 38

4.10 Regression results ($N = 169$) for verbal, nonverbal and combining both cues using RF ($p < 0.05$ for all). 39

4.11 List of top 20 features from RF regression model ($N = 169$) for *Overall Impression* (left:$1-10$; right:$11-20$) 40

4.12 Inferring of principal components (PC) using random forest (RF) with verbal, nonverbal and their combining as predictors ($p < 0.05$). 40

5.1 Reception desk dataset: $ICC(2,k)$ & descriptive statistics for impressions of skills and performance, attractiveness & personality traits. 46

5.2 Nonverbal features extracted from the reception desk data. 46

5.3 Correlation between Big-5 personality trait impressions and *Performance Impressions* (N = 169; $^{*}p < 0.001$, $^{\dagger}p < 0.01$, $^{**}p < 0.05$). Entries without p-value symbol are not statistically significant. 47
5.4 Pearson correlation between nonverbal cues and performance and skill impressions N = 163;$^{*}p < 0.001$,$^{\dagger}p < 0.01$, $^{**}p < 0.05$. 48
5.5 Range of Pearsons correlation between eye gaze, facial expressions and social variables in the reception desk (N = 153) dataset. Due to nonfrontal face in this setting, the N is different from Tabl 6.7. $^{***}p < 0.001$;$^{**}p < 0.01$;$^{*}p < 0.05$. . . . 49
5.6 Correlation between *Attractiveness* attributes and *Performance Impression* (N = 163;$^{*}p < 0.001$,$^{\dagger}p < 0.01$, $^{**}p < 0.05$). Entries without p-value symbol are not statistically significant. 50
5.7 Pearson correlation between nonverbal cues and performance and skill impressions N = 163;$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ 51
5.8 Best inference performance results using NVB cues, personality traits (Big-5) impressions, Attractiveness impressions and various combination of impressions and NVB. All results were significant with $p < 0.05$ (N = 169). Best performing model is indicated by *(RF); ** (Ridge) 52
5.9 Regression analysis using eye gaze, facial expressions and combination of features for desk (N = 153) setting. 53
5.10 Inference results using verbal cues, fusion of verbal cues with NVB cues and personality traits (Big-5) impressions. All results were significant with $p < 0.05$ (N = 169). Best performance is obtained using RF model. 55

6.1 List of perceived variables manually annotated for both situations, along with their $ICC(2,k)$ and means. 60
6.2 Pearson's correlation between perceived variables from interview (I) and reception desk (D) situations (N = 169). All of them are significant with $p < 0.001$ . . 63
6.3 List of predictors used in regression experiments obtained from job interview (I) and reception desk (D) interactions. 64
6.4 Summary of experiments and the best regression performance (R^2) achieved. All results are significant with $p < 0.05$. 65
6.5 Top 20 variable importance in the RF for *Exp1c*. All measures of importance indicated in the *Rank* column are scaled to have a maximum value of 100. . . . 66
6.6 Pearson's correlation between perceived variables of desk and nonverbal cues displayed during job interviews (N = 169). All features are significant $^{***}p < 0.001$,$^{**}p < 0.01$, $^{*}p < 0.05$ 67
6.7 Selected Pearson's correlation coefficient for perceived hirability (Rating(I)) and perceived performance (Rating(D)) across the two situations (N = 169). $^{**}p < 0.01$, $^{*}p < 0.05$ 68
6.8 Summary of experiments and best regression performance (R^2) of desk perceived variables achieved. All results are significant with $p < 0.05$. 69

6.9 Summary of experiments with linguistic content and the best inference performance achieved. All results are significant with $p < 0.05$. 70

7.1 Inter-rater agreement ($ICC(1,k)$) and descriptive statistics for crowd-sourced annotations with each video being annotated by 5 raters. Total number of videos are 939 (Source: Nguyen et al. [116]) 79
7.2 Results of the inference task using the random forest algorithm (N=292) and features extracted from the manually transcribed text corpus as predictors. . . 85
7.3 Results of the inference task using the random forest algorithm. (N=292). The manually transcribed (Manual) and automatically transcribed (ASR-2min) text corpus were used as predictors. 86
7.4 Results of the inference task using the random forest algorithm. Both the automatically transcribed text corpus, ASR-2min and ASR-Full were used as predictors. (N=292) 88
7.5 List of questions in the self reported pre- and post-questionnaires rated on a Likert scale 94
7.6 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. 97

1 Introduction

Interpersonal communication represents the means through which we initiate, negotiate, and maintain human relationships [101]. Hence, interpersonal skills are paramount in the context of workplaces and are critical in certain sectors like hospitality, sales, and marketing. Interpersonal communications also has implications in the formation of first impressions. First impressions can be defined as the mental image one forms about something or someone after a first encounter. Literature in psychology has shown that people can make accurate inferences about others, even after a short duration of interaction [7, 4, 9]. In professional spheres, these initial judgments can influence critical outcomes, such as being hired or promoted. It has also been demonstrated in the literature that as a major component of interpersonal communication, nonverbal behavior (NVB) contributes towards the formation of first impressions [87, 60].

The advent of inexpensive, ubiquitous and unobtrusive sensors in the last decade has lead to increasing ease of sensing and interpreting social signals conveyed by users up to a certain level [162, 58, 131]. The development of computational models has led to the incorporation of technology to sense and analyze interpersonal behavior. This domain, known as *social sensing* [50], utilizes multiple modalities (audio, video and mobile sensors) to automatically capture and analyze human behavior, and interpret the social signals they convey. Literature in computing has shown the feasibility of using such methods to automatically infer various variables of interest in dyadic and group interaction including personality [15, 18, 132], emergent leadership [140, 143], dominance [79, 76] or hirability [114, 112, 28].

1.1 Motivation

In the hospitality industry, interpersonal skills displayed by service employees (e.g. reception desk assistants) during dyadic interactions with customers (commonly referred to as service encounters) form a critical part of customer experience at an establishment [157]. Customers perceive and evaluate the employee's attitudes, professional, social, and communication skills during these encounters, and form impressions of both the employee and the organization

based on these short interactions [103]. The implications of interpersonal communication on customer and perceived quality of service (QoS) have been previously investigated in hospitality and management literature [55, 157].

Given the importance of interpersonal communication in hospitality, this thesis, using advances in ubiquitous sensing, audio-visual processing, and machine learning, investigates the automatic recognition of first impressions in two dyadic interaction settings in hospitality. Specifically, we identify two important situations, namely (a) employment interviews and (b) interactions with customers at a reception desk, to understand the formation of first impressions by observers of the interaction, and the subsequent automated inference tasks. The choice of the job interview setting was motivated by its ubiquity in the process of selecting new personnel. The choice of the interaction with a client at a reception desk was motivated by the fact that it constitutes a standard type of interaction in hospitality. Another reason for the choice of this setting is the need for good interpersonal communication with customers, during service failures to maintaining good impressions.

This thesis also investigates the role of a setting in the formation of first impressions. This is important as people behave differently in different situations, as person and situation are intricately entwined. Also known as "person-situation debate" or "person-situation-behavior triad", this area has been a topic of research in social sciences for decades [83]. Yet, until the advent of ubiquitous computing technologies, it had been difficult to objectively quantify behavior in multiple person-situation cases, due to lack of access (both direct and unobtrusive) to interactions across situations [51, 100]. To the best of our knowledge, there have been no studies to automatically infer first impressions in two hospitality situations individually and across two situations in the context of workplaces.

Another area of focus in this thesis is the evaluation of a real-time feedback system for dyadic interactions. Advances in ubiquitous and wearable computing with smart interfaces have enabled the development of behavioral training systems [64, 119] and other applications in the classroom (like physics experiments [165]). Behavioral training, both real-time and offline, has been investigated in various domains including public speaking [37, 158], group interactions [41, 10, 86, 156], and interviews [72]. In this thesis, we design and implement an automatic, unobtrusive, real-time, conversational behavior awareness system for young sales apprentices to make them aware of their nonverbal behavior while interacting with a customer. The choice of nonverbal behavior was guided by the results of our investigation in dyadic workplace interactions [87, 47, 114, 108] and the low computational power of Google Glass device. The system can provide real-time feedback during a face-to-face interaction at workplaces to increase self-awareness of conversational behavior and does not impair the quality of interaction.

This thesis also investigates the connections between verbal content and first impressions in noisy, "in-the-wild" video. Most existing works have investigated first impressions from the perspective of nonverbal behavior with verbal content receiving little interest. This is mostly

due to manual transcriptions of social interactions being a long, tedious and expensive process. The advent of deep convolution neural networks has enabled automatic speech recognition systems to reach near-human performance levels, as well as, improved representations of text documents. In this thesis, we leverage these recent advances to investigate the relationship between verbal content and the formation of first impressions in conversational video resumes, which are short video recorded messages where job-seekers present themselves to potential employees. To this end, we use 292 noisy, "in-the-wild" video resumes from YouTube, develop a computational framework to automatically extract various representations of verbal content, and evaluate them in a regression task.

1.2 Objectives

The objective of this thesis was to build upon literature, to employ ubiquitous multimodal sensors to capture and quantify the interpersonal interactions, their relations to first impressions and to provide real-time unobtrusive feedback for improved behavioral awareness in professional settings. It must be noted here that the term first impression also includes impression management, which is defined in literature as conscious or subconscious process in which people attempt to influence the perceptions of others. To this end, we used a living lab approach to acquire a novel corpus consisting of human behavioral cues from the same sample of population in two different workplace scenarios.Based on this data, we developed computational methods to automatically analyze the relationship between verbal and nonverbal behavior, first impressions, hirability and other related variables. We then used these insights to explore two possible applications, one about real-time feedback, and one about online video resume

The specific objectives of this thesis are the following:

1. Automatic characterization of first impressions across two diverse hospitality settings.
    - Design and implementation of a behavioral training procedure for hospitality students to help improve the first impressions they make in their professional sphere.
    - Investigation of automatically extracted nonverbal cues that are predictive of first impression in job interviews and reception desk.
    - Investigation of the role of verbal content (extracted using natural language processing) in the inference of soft skills impressions in the two workplace settings.
    - Assessment of what variables are predictive of (both verbal and nonverbal cues) first impressions are situation-dependent.
2. Investigation of two additional application scenarios of nonverbal and verbal automatic analysis.
    - Assessment of various NLP based representation of verbal content and to infer first impressions in video resumes.
    - Investigate the effect of automatic speech recognition on inference performance

compared to manual transcription.
- Assess the usefulness and ease of feedback incorporation in modulation of behavior to positively affect first impressions.

1.3 Contributions

The contributions of this thesis are the following:

1. **Data collection:** We contributed to the collection of a new dataset resulting from a training framework for students in an international hospitality school. We described a number of challenges associated with implementing a behavioral training procedure for hospitality students in order to improve the first impressions that others form about them. The framework consisted of two scenarios that were relevant for their future careers, namely job interviews and reception desk situations. We collected a new corpus of 169 simulated job interviews and reception desk interactions (338 interactions), which to our knowledge constitutes one of the largest academic datasets of work-related dyadic interactions. This dataset was recorded with multiple modalities. Using the state-of-the-art audio and video processing methods, we extracted a number of behavioral cues from both the settings. The choice of these behavioral cues were backed up by literature in psychology, hospitality and computing. The variable of interest in both of these situations was obtained by manual annotations.

2. **First impressions in job interviews :** We analyzed the 169 job interview videos to understand the relationship between automatically extracted nonverbal cues, verbal cues and various perceived social variables in a correlation analysis. Inferring *Overall Impression* in a regression task gave $R^2 = 0.32$, thus extending the results found in [114]. We found gender differences that confirmed previous findings in psychology [90] We transcribed interview video and analyze the linguistic content using off-the-shelf software. Our results indicated low predictive power of verbal cues for *Overall Impression* ($R^2 = 0.11$). Fusion of verbal and nonverbal cues improved inference performance ($R^2 = 0.34$). To understand the structure of the soft skill impressions, we conduct a principal component analysis. The use of principal components revealed a major component associated with overall positive and negative impressions that when used as labels in a supervised learning results in a regression performance of $R^2 = 0.41$.

3. **First Impressions at Reception Desk :** We analyzed the relationship between automatically extracted audio-visual nonverbal cues and performance and skill impressions via correlation analysis and a regression task, and obtained $R^2 = 0.30$. We observed that Big-5 trait impressions achieve performance of $R^2 = 0.35$ for job performance impressions. An interesting result obtained was the low connections between judgments of attractiveness and performance and skill impressions. Best inference results were obtained by fusing NVB cues and Big-5 impression results with $R^2 = 0.37$. This results

have implications for employees and managers in hospitality and could also facilitate personalized training for employees to improve their nonverbal behavior in service encounters.

4. **Cross-Situation Analysis in Hospitality:** We investigated human behavior across two situations (job interview and reception desk) in a data corpus of 338 dyadic interactions, with the objective of inferring performance on the job. We first conducted a cross-situation Pearson's correlation analysis between perceived hirability and soft skills at job interviews, and perceived performance and soft skills at the reception desk. We observed that perceived variables in job interviews are moderate indicators of perceived performance and soft skills on the job. Our best inference model obtained R^2 = 0.40, using a fusion of nonverbal cues extracted from the reception desk interaction and the human-rated interview scores. An interesting observation was that some behavioral cues (greater speaking turn duration and head nods) were positively correlated to higher ratings for all perceived variables independent of the situation. Using a fusion of LIWC and Doc2Vec features to represent verbal content, the best inference performance of R^2 = 0.25 was achieved for perceived performance.

5. **Verbal Content and First Impressions in Video Resumes:** We analyzed the relationship between verbal content and the formation of first impressions in 292 noisy, "in-the-wild" video resumes from YouTube. Towards this, we leveraged existing NLP to investigate the various text representations methods. Video resumes were transcribed both manually and automatically (using cloud-based Google Speech API). We then extracted various representations of verbal content including LIWC, Doc2Vec, Word2Vec and GloVe.

6. **Real-time Behavioral Feedback System:** We evaluated a conversational behavior awareness tool using Google Glass. The goal of the system was to provide real-time feedback to young sales apprentices about the amount of time they talked during a sales interaction with a client. The effectiveness and unobtrusiveness of the system was evaluated by conducting a pilot study involving 15 apprentices (aged between 16-20). Overall, participants found the real-time feedback system fun, little distracting and useful. Furthermore, analysis of the manual coding of the interaction videos showed no negatively influence of the wearable sensing and real-time feedback on the dyadic social interaction.

1.4 Outcomes

We believe that the insights from our work have implications for hospitality and other customer-facing domains, organizational psychology and computing. For psychologists, our work provides insights on what verbal and nonverbal cues play an important role in the formation of first impressions across multiple situations in workplaces. Also, our work shows the feasibility of unobtrusive, multimodal sensing and objective quantification of behavior for assessing candidates in dyadic hospitality interactions. This work also has wider implications for em-

ployees and managers in hospitality, by providing an understanding of not only the employee's performance via nonverbal behavior, but also of the implications for customer perceptions of service encounters. In computing, our research has the potential to enable several applications. For example, the automatic approach could facilitate personalized training for employees to improve their nonverbal behavior in service encounters. The results of this thesis is also potentially important to socially challenged individuals to express and/or perceive nonverbal communication. Overall, understanding differences in behavior in two diverse settings coupled with cross-situation analysis and the information they convey is important towards building ubiquitous computational devices capable of sensing and responding unobtrusively [121, 164].

1.5 Organization of this Thesis

The rest of this thesis is structured as follows. In Chapter 2, we present the existing literature from organization psychology, hospitality management, and computing. In Chapter 3, we describe the context and challenges in the design and development of a living lab, and the data collection process. This chapter further outlines the manual annotations conducted, the various multimodal nonverbal behavioral cues and natural language processing based text representations extracted from the data corpus. Chapter 4 investigates the job interview part of the corpus, while Chapter 5 investigates the reception desk videos. In Chapter 6, we investigate automatic inference of desk performance using cross-situational behavioral cues. In Chapter 7, we present two additional applications of our studies (a) understand behavior in a real-world data source and (b) develop a system to provide real-time feedback for user behavioral awareness. Chapter 8 concludes this thesis by discussing the limitations of our work and potential directions for future work.

1.6 Publications

This thesis is a compilation of works published in one international journal and six conference/workshop proceedings.

Journal Paper

Muralidhar, Skanda and Mast, Marianne Schmid and Gatica-Perez, Daniel. A Tale of Two Interactions: Inferring Performance in Hospitality Encounters from Cross-Situation Social Sensing. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2018.

Conference and Workshop Papers

Muralidhar, Skanda and Siegfried, Rémy and Odobez, Jean-Marc and Gatica-Perez, Daniel. Facing Employers and Customers: What Do Gaze and Expressions Tell About Soft Skills? In: *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*,

2018.

Muralidhar, Skanda and Nguyen, Laurent Son and Gatica-Perez, Daniel. Words Worth: Verbal Content and Hirability Impressions in YouTube Video Resumes. In: *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2018.

Muralidhar, Skanda and Gatica-Perez, Daniel. Examining Linguistic Content and Skill Impression Structure for Job Interview Analytics in Hospitality. In: *Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia*, 2017.

Muralidhar, Skanda and Mast, Marianne Schmid and Gatica-Perez, Daniel. How May I Help You? Behavior and Impressions in Hospitality Service Encounters. In: *Proceedings of the 19th ACM on International Conference on Multimodal Interaction*, 2017.

Muralidhar, Skanda and Nguyen, Laurent Son and Costa, Jean and Gatica-Perez, Daniel. Dites-Moi: Wearable Feedback on Conversational Behavior. In: *Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia*, 2016.

Muralidhar, Skanda and Nguyen, Laurent Son and Frauendorfer, Denise and Odobez, Jean-Marc and Mast, Marianne Schmid and Gatica-Perez, Daniel. Training on the Job: Behavioral Analysis of Job Interviews in Hospitality. In: *Proceedings of the 18th ACM on International Conference on Multimodal Interaction*, 2016.

Finnerty, Ailbhe and **Muralidhar, Skanda** and Nguyen, Laurent Son and Pianesi, Fabio and Gatica-Perez, Daniel. Stressful First Impressions in Job Interviews. In: *Proc. of the 18th ACM on Int Conf. on Multimodal Interaction*, 2016.

2 Related Work

Understanding the different social signals conveyed by human behavior (verbal and nonverbal) is an important area in psychology and human computer interaction (HCI). Psychology literature reports that nonverbal behavior component of human behavior is an important contributor to the formation of first impressions [87, 145]. First impressions are defined as a "mental image formed of a person when met for the first time" [9]. The formation of accurate impressions from a small duration of interaction ("Thin slices") has validation in psychology [7] in various settings . This thesis investigates the interconnections between nonverbal behavior, verbal content, and first impressions in two different workplace situations (interviewing for a job and performing at the job) and contributes to the literature.

Most research in social psychology, hospitality, and marketing relies on manual annotations of nonverbal behavior, making it labor-intensive and difficult to scale with respect to large number of users and different scenarios. The advent of ubiquitous sensors combined with improved perceptual techniques have enabled the automatic analysis of human interactions [58, 130].

In this section, we outline the related works in psychology, hospitality and social computing. In this chapter, we first present the literature in both hospitality and psychology (Section 2.1). We then present the literature in social computing (Section 2.2).

2.1 Literature in Psychology & Hospitality

2.1.1 Job Interviews

Researchers in organizational psychology have studied job interviews for decades, uncovering statistical relationships between nonverbal behavior, personality, hireability, and job performance. Regardless of the setting, the impact of nonverbal behavior on perceiver's impressions and judgments has been established [145]. In the context of job interviews, the applicant's nonverbal behavior was shown to influence the hiring decision of the recruiter. For example,

Imada and Hankel showed that high amount of eye contact, smiling and other nonverbal behavior had a significant effect on the outcome of the interview [77]. Furthermore, successful applicants were found to make more direct eye contact, produce more facial expressions, smile and nod more than applicants who were rejected [47]. McGovern and Tinsley reported that applicants with loud and modulating voice, extended eye contact, fluent speech, and expressive face were more likely to be hired than the applicants who did not show such behavior [99]. Along the same lines, powerless speech (i.e., speech punctuated with hesitation and lacking conviction) had a negative effect on the impression ratings compared to applicants with speech disorders like stuttering or lisping [45]. Until recently, research in social psychology relied on manual annotations of nonverbal behavior, which is labor-intensive and difficult to scale with respect to either large number of users or different scenarios.

2.1.2 Perceived Performance

First impressions are defined as formation of a mental image of a person when met for the first time [9]. Research in psychology has revealed that nonverbal behavior is an important component in the formation of first impressions [87, 145] and that even a short interaction ("Thin slices") is enough to form first impressions [7]. Specifically in the workplace setting, thin slices of nonverbal behavior have shown to be predictive of job interviews [115], evaluation of sales job performance [6], and employee-customer interaction [11]. Regarding assessment of performance, Ambady et al. showed that end of semester ratings of 13 university teachers as rated by students could be predicted based on judgments of personality characteristics from 10-second clips [8]. The same authors showed the predictive validity of thin slice judgments on the performance of 12 sales managers using 30-sec audio clips [6].

The influence of physical attractiveness on impressions has been investigated in psychology. Ahearne et al. reported a positive effect of physical attractiveness on sales performance of 339 pharmaceutical sales representatives [2]. Here, attractiveness was rated by physicians based on their personal interaction with sales representatives. Similarly, Hamermesh et al. reported that college professors perceived as physically attractive were evaluated higher by students [66]. In this study, six undergraduate students (3 females) rated perceived attractiveness of 94 professors using photos posted on department websites. Magnus et al. investigated the physical attractiveness of service workers in two settings (bookstores and airline travel) and reported the positive impact of physical attractiveness on customer satisfaction in both settings [151]. In this study, attractiveness was rated based on photos of the service workers.

In the context of hospitality, sales and marketing, and management, the relationship between impression formation and nonverbal behavior has also been acknowledged [22, 157]. Many aspects of nonverbal behavior including gestures, smiles, touch, and prosody as well as other attributes like physical attractiveness have been explored. Gabbott and Hogg, using a study conducted using a video recording of an actress playing the role of a reception desk assistant helping a customer to check-in, showed that nonverbal communication impacts the customer

perception of QoS [54]. Furthermore, the study also showed the effect of perceived attractiveness on satisfaction of service. Kang and Hyun investigated the effect of communication styles on customer-oriented service in a study consisting of 527 luxury restaurant patrons [82]. In this study, the authors emailed a questionnaire to participants based on their visit to a luxury restaurant. They reported that employees who smiled, nodded and maintained eye contact with the clients, and spoke with high energy and tone of voice with fewer short utterances were positively correlated to customer satisfaction.

Jung and Yoon, in a study consisted of 333 customers, investigated the role of nonverbal behavior in customer satisfaction at a family restaurant in South Korea [81]. The authors reported a positive correlation between visual nonverbal cues (gestures, head nods) and customer satisfaction ($r = 0.42$; $p < 0.01$), and between paralingual nonverbal cues and customer satisfaction ($r = 0.33$; $p < 0.01$).

The relationship between personality and job performance impressions was investigated in [12]. The authors conducted a meta-analysis and reported correlations between Big-5 personality traits and job performance. Specifically, they found that *Conscientiousness* was correlated to all occupational groups in the study, while *Extraversion* was found to be a valid predictor for jobs that require social interactions, like managers.

The effect of physical attractiveness has been a subject of interest in the marketing and service industry. In hospitality, it was shown that tips received by female waitresses from male customers were positively related to service providers wearing makeup [78] and certain colored clothes [63]. Both these studies were conducted in the field, and attractiveness was rated based on physical appearance during the interaction in a restaurant setting. Similarly, Luoh et al. reported that customer's perceptions of service quality were enhanced by attractive service providers compared to those of average appearance [94]. A "Beauty is beastly" effect was reported in a study which found that physical attractiveness for women was detrimental in employment contexts considered to be masculine (e.g managers, director of security etc) [80]. In both studies [80, 94], attractiveness was rated based on photographs controlled for background, age, and posture.

2.1.3 Cross Situation Analysis in Literature

Due to the difficulty in obtaining direct behavioral measurement across multiple situations, it has been traditionally challenging to quantify behavior in "person-situation" interactions [51]. There are few works in psychology that have investigated impressions and behavior across multiple situations, especially those investigating behavior displayed during job interviews with performance on the job. Motowildo et al. investigated the connection between aural and visual cues displayed in a structured interview of 40 managers from a utility company with the performance ratings of these managers by their supervisors [104]. The recorded interviews were rated by 194 undergraduate students on the same scale. The study reported a correlation of $r = 0.36$ between the student ratings and supervisors ratings of performance.

This work was extended by DeGroot et al. who investigated various nonverbal cues and their correlation to performance ratings [40]. This work used a dataset of 110 managers from a news-publishing company, and reported that vocal cues (pitch, pitch variability, speech rate, and pauses) correlated to performance ratings ($r = 0.20$; $p < 0.05$). The study also reported that visual cues (physical attractiveness, smiling, gaze, hand movement, and body orientation) had low correlation to performance ratings ($r = 0.14$; $p < 0.05$).

All of the above research in the domain of social sciences including psychology, marketing and hospitality has so far relied on manual behavior coding. This process is expensive and laborious, and hence it is difficult to investigate many features or multiple situations, making such studies rare in the literature [51]. This situation has advanced through ubiquitous technologies.

2.2 Literature in Computing

2.2.1 Job Interviews

The advent of inexpensive and unobtrusive sensors combined with improved perceptual techniques have enabled the automatic analysis of human interactions. This domain is multidisciplinary and involves speech processing, computer vision, machine learning, and ubiquitous computing. Early works investigated the use of automatically extracted nonverbal cues to predict social constructs as diverse as dominance, leadership, or personality traits in small group interactions [59]. In a context similar to job interviews, Curhan et al. [36] investigated the relationship between audio cues and social outcomes in dyadic job negotiations. Batrinca et al. [15] used a computational approach to predict Big-Five personality traits in self-presentations where participants had to introduce themselves in front of a computer, in a manner similar to job interviews, but without the presence of an interviewer. Nguyen et al. [114] addressed the problem of automatically analyzing employment interviews. This work used automatically extracted nonverbal cues (speaking turns, prosody, head nods, visual activity) to infer five types of hireability variables in a dataset of 62 real job interviews. Further research also examined the relationships between body activity, personality and hireability using a mixture of automatically and manually extracted features [117]. Naim et al. [112] extended these works by analyzing a dataset of 138 simulated job interviews from 69 internship seeking students, where they extracted cues related to facial expressions, verbal content, and prosody to predict several social variables (e.g., hiring recommendation, engagement, friendliness) and perceived behaviors (e.g., smile, eye contact, speaking rate).

Existing literature in psychology indicates that people can improve their chances of getting hired in a job interview by practicing both their verbal and nonverbal communication [71]. The recent advances in wearable devices with smart interfaces have enabled a new range of possibilities for behavioral training [119]. In the context of public speaking, Google Glass has been used as a head-mounted display system to provide real-time feedback on a presenter's

posture, openness, body energy, and speech rate sensed using Kinect and external microphone data [37]. In addition to displaying information, Google Glass was also used as an audio sensor to provide automatic real-time feedback on a speaker's speaking rate and energy [158]; however, the data was processed on an external server. These systems have been evaluated on relatively small cohorts ($N \in [15,30]$) consisting of computer science students, and to this day several important questions remain unanswered, such as the implementation of such systems in realistic settings, or the social acceptability of Google Glass. In the context of job interviews, MACH [72] was developed to train social skills and consists of a virtual agent able to read behavioral cues (facial expressions, speech, and prosody) produced by a participant. Additionally, the system provides summary feedback on various nonverbal cues (smiles, head nods, pauses, etc.), as well as what authors called focused feedback, which consists of visualization of certain behaviors over time along the recorded video. The system was tested in a job interview scenario on a cohort of 60 students (plus 30 as control group). While results showed that the group who was given feedback improved their overall interview performance, little is known on how to implement a training procedure that can be used by individuals and organizations, such as schools or employment agencies to systematically benefit from it.

2.2.2 Perceived Performance

Literature in social computing has validated the viability to integrate nonverbal behavior extracted using ubiquitous sensors and machine learning algorithms to infer various constructs like interview ratings [114, 112], negotiation outcomes [36], and Big-5 personality [15] to promising levels [59].

Batrinca et al. used an approach to predict Big-5 traits from self-presentation questions, where participants introduced themselves in front of a computer, similar to job interviews, but without the presence of an interviewer [15]. Nguyen et al. used automatically extracted nonverbal cues (speaking turns, prosody, head nods, visual activity) from both applicant and interviewer to infer five hirability variables in a dataset consisting of 62 real job interviews [114]. Naim et al. extended these works by analyzing a dataset of 138 simulated job interviews from internship-seeking students [112]. The authors extracted cues related to facial expressions, verbal content, and prosody to predict several variables (hiring recommendation, engagement, friendliness). Chen et al. developed a standardized video interview protocol along with human ratings, which focused on verbal content, personality, and holistic judgment [28]. The authors using "visual words" as feature extraction method, automatically learned from video analysis outputs, and the Doc2Vec paradigm achieved a correlation of 0.42 between machine-predicted scores and human-rated scores.

Biel et al. studied effects of physical attractiveness in a study focusing on modeling different facets of YouTube vloggers [19]. Using 442 vlogs rated for two dimensions of physical attractiveness, and three dimensions of non-physical attractiveness, they reported significant

positive correlations between judgments of attractiveness and two Big-5 traits (Extraversion and Agreeableness). Attractiveness was rated on 1-min vlogs by Amazon Mechanical Turk workers.

In the context of performance, Curhan et al. investigated the relationship between audio cues and negotiation outcomes in dyadic job negotiations [36]. Performance was measured as the compensation package that could be negotiated. The authors reported that voice activity levels, prosodic emphasis, and vocal mirroring explained up to 30% of the variance. Madan et al reported the validity of audio nonverbal cues in predicting the performance of male participants in a speed dating setup [95]. In a study consisting of 57 five-minute sessions, the authors reported positive correlation between a female 'liking' a male participant and aggregated male and female speech features ($r = 0.67$; $p < 0.05$). In this setup, performance was measured as the number of likes received from female participants. Lepri et al. investigated the use of nonverbal behavior for inference of individual performance in a group task [92]. Using audio and visual features extracted from the Mission Survival 2 corpus, they were able to classify binary levels of performance with accuracy up to 50%. Raducanu et al. used a dataset collected from the reality show "The Apprentice" to predict the person who will be fired [133]. Using speaking turn features, the method predicted the candidate to be fired (i.e the one with worst perceived performance) with an accuracy of 92%.

The existing literature in psychology and social computing demonstrates the feasibility of predicting first impressions up to a certain degree using nonverbal and verbal cues including other constructs like attractiveness. In this thesis, we investigate the interplay between nonverbal and verbal behavioral cues, and impressions of attractiveness, personality traits, and performance in two workplace settings. To the best of our knowledge, there have been no studies to automatically infer first impressions in two hospitality situations individually and across two situations in the context of workplaces. Our work, therefore, could have wider implications for managers and training students in the fields of customer service and hospitality.

3 Data Collection & Feature Extraction

The understanding of both human perception and automatic recognition of first impressions in two different workplace situations is a key objective of this thesis. To address this problem, we collect a novel dataset comprising of 169 participants in two diverse situations: job interviews and hotel reception desk. The collection of this dataset, known as the UBIMPRESSED dataset, is additionally relevant given the lack of datasets consisting of the same participants in two different workplace situations. This dataset was collected as part of the Swiss National Science Foundation UBIMPRESSED (Ubiquitous First Impressions and Ubiquitous Awareness) project, and is a collaboration between Idiap Research Institute (Prof. Daniel Gatica-Perez), University of Lausanne (Prof. Marianne Schmid-Mast), Cornell University (Prof. Tanzeem Choudhury), and Vatel International Hospitality Management School. The work described in this Chapter was a collaboration with Dr. Laurent Nguyen (Idiap), Dr. Denise Frauendorfer (UNIL), and was supported by a number of research assistants.

In this chapter, we present the data collection, annotation and feature extraction details. Specifically, Section 3.1 outlines the context and challenges during this of this data collection phase, Section 3.2 describes the experimental design and technical setup. Section 3.3 presents the details of the manual annotation of a set of social variables of interest, while Section 3.4 details the manual transcription. Finally, Section 3.5 describes the various behavioral cues captured and features extracted along with the literature supporting our choices.

The material of the chapter was originally published in [108].

3.1 Context and Challenges

One of the main objectives of this work was to design and implement a behavioral training framework for students in an international hospitality school offering bachelor and master degrees for English and French-speaking students. The envisioned framework involves students practicing some of their regular work tasks in realistic conditions. We faced many challenges associated with the real-world implementation of a behavioral training program.

The hospitality management programs are immersive with students taking classes and conducting practical work in a real hotel, where they rotate among the different services (kitchen, reception, bar, etc.). First, classes had weekly-rotating shifts of practical work and courses. For practical weeks, students did not know their schedules more than two days in advance. Classes did not start semesters at the same period of the year. These factors inherent to the school made the planning of the data collection complex, both at the semester and week levels, requiring a high level of flexibility from our side. Second, as in any bachelor or master level program, students were busy with their usual curricular activities (course assignments, projects, mid-terms, and finals) that took a significant amount of their time. Third, we also faced some challenges related to the relatively young age of the students: some of them were not 100% reliable and did not show up at scheduled sessions, did not reply to emails/SMS, or dropped out during the course of the study. Additionally, although the benefits of participating in the training program was clear to the majority of students, some felt the investment in time was too high to participate.

To address these challenges, we built a living lab located in the same building as the hospitality management school. We had three evenings per week where the lab was open and students could register up to 24 hours in advance for a training session, with a maximal capacity of 12 student-sessions per week. To avoid overloading the student's schedule, we made efforts to make the training program as time-efficient as possible; in total, the complete procedure for each student took 4 hours distributed over 4 weeks on average. Additionally, we targeted students who were starting their semester, because it corresponded to the time where their school-related workload was the lowest.

Multiple modalities were used to advertise the program. Subscription sheets summarizing the study were placed at the student help-desk; 10-minute class presentations were given to each class within the first two weeks following the start of their semesters, where we listed the benefits for them to participate in the training program; e-mails by the academic directors of the school were sent to students; last, some professors advertised the study during their class. Furthermore, we incentivised the students to participate by offering the equivalent of USD 50 upon the completion of the program. Participation in the program was voluntary.

3.2 Overall Design

The behavioral training program was designed to be beneficial for the students. To this end, we identified two important situations in the context of hospitality where behavior plays an important role: employment interviews and interactions with customers at a reception desk. The choice of job interview was motivated by its ubiquity in the process of selecting new personnel; furthermore, we believed that a behavioral training on interviews could be beneficial for students in the relatively short-term to get hired for an internship (which students need to complete as part of their degree requirements) or to land their first job. The choice of the interaction with a client at a reception desk was motivated by the fact

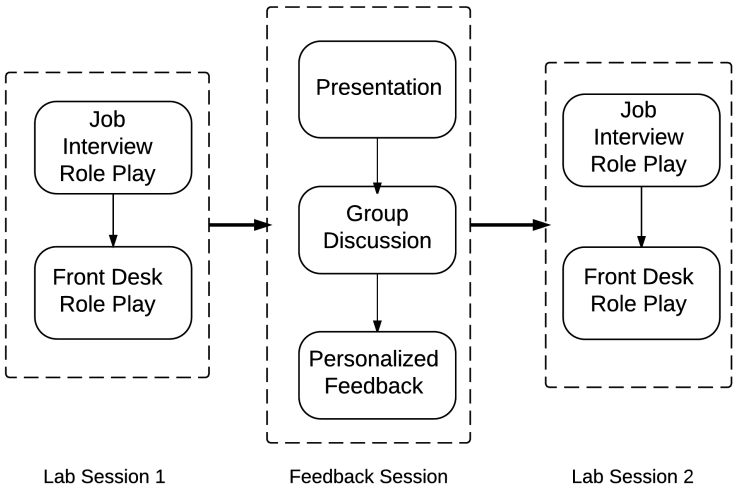


Figure 3.1 – Illustration of the design of the behavioral training procedure.

that it constitutes a standard type of interaction in hospitality; furthermore, beyond the reception desk, hospitality professionals have to be able to communicate with (possibly angry) customers, regardless of the setting. In addition to the two role-plays, an interactive feedback session was included in the training process. This feedback session was held in groups of 3 to 8 students, in which they were given a presentation on first impressions and behavior, watched and discussed video snippets of their recorded interview and desk role-plays, and received written personalized feedback by professionals in human resources and hospitality. Specifically, the procedure included two laboratory sessions, with a feedback session held in-between. Each laboratory session included two role-plays: one job interview and one front desk interaction. Figure 3.1 displays an overview of the behavioral training framework.

(a)

Interview Questions

1. Short self-presentation.
2. Motivation for working in the service industry.
3. Past experience requiring attention to details.
4. Past experience where stress was correctly managed.
5. Past experience where adaptability was required.
6. Past experience of calm and tact under stress.
7. Strong/weak points about self.

(b)

Figure 3.2 – (a) Snapshot of the job interview situation collected as part of the behavioral training program; (b) Questions asked during job interview where the participant is role playing an applicant for an internship in a high-end hotel.

3.2.1 Lab Sessions

Each laboratory session consisted of a job interview and a reception desk role play. The scenarios of lab sessions 1 and 2 were identical to the exception of the reception desk scenario which was slightly modified.

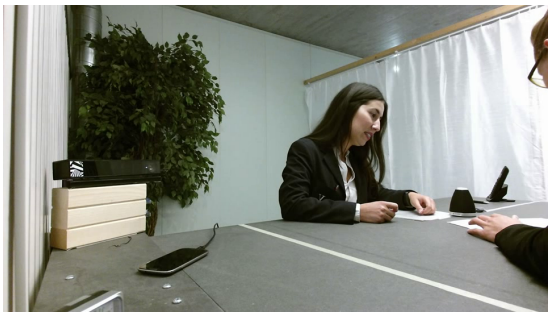
Both the scenarios were recorded using two Kinect v2 devices, one for each protagonist in the interaction. These sensors record standard RGB color and depth data at 30 frames per second, with a spatial resolution of 1920 × 1080 for RGB and 512 × 424 for depth. Additionally, a Kinect v1 device was placed on top of the reception desk to provide a bird's view of the interaction. For audio, we used a Microcone [1] device which is an array of microphones that automatically performs speaker segmentation based on sound source localization, in addition to recording audio at 48kHz. Cross-sensor synchronization was obtained by manually adjusting the delay between video data with respect to audio data.

Job Interview

The job interview situation consists of the participant applying for an internship at a high-end hotel (Figure 3.2b). A structured interview process, with each interview following the same sequence of questions, was employed. This process has been shown in psychology to be among the best tools to compare interviewees and select applicants [74]. The interviews were conducted by a pool of seven trained research assistants who were university students in organizational psychology and business.

Reception Desk

The reception desk situation consists of a role play between a receptionist (the participant) and a client (one of the research assistants). The participants were informed of the situation before starting (Figure 3.3b) but are unaware of the client's exact reaction. The aim of the situation was to assess participants' ability to handle an unfriendly client in the best possible manner.



(a)

The participant is a reception desk intern at a high-end hotel. The manager is unavailable & the participant has to handle all customer interactions. A client comes to check out. The initial interaction is friendly. The customer's attitude becomes unfriendly after the bill is recieved and starts complaining about issues faced during the stay. The participant needs to handle the situation and come to a resolution that is acceptable to the client.

(b)

Figure 3.3 – (a) Snapshot of the hotel reception desk situation collected as part of the behavioral training program; (b) Details of the desk situation where participants have to handle an unhappy client.

The scenario was slightly modified in the second lab session to reduce its predictability. In this session, the participants interact with a new client (i.e., a different research assistant) who changes her/his attitude even before receiving the bill by complaining about a bad restaurant recommendation by a previous receptionist.

We believe that the reception desk scenario is an interesting addition to the job interviews, as it constitutes a situation where the students perform in a work-like environment. In other words, this scenario could provide an assessment of how well the students perform in a job situation, which could enable us to study the relationship between job interviews and the performance at work.

3.2.2 Feedback Session

After the first lab session, students participated in a feedback session in groups of three to eight people. First, participants were given a 20-minute presentation on nonverbal behavior and its effect on the first impressions made on others. The presentation was prepared by a researcher in organizational psychology and was given by one of our research assistants.

Then, video snippets of the interactions recorded during the first lab session were watched and discussed by the group of students. Participants were instructed to give constructive comments about several strong points and aspects that could be improved. At least one research assistant was present to moderate the discussion, which was never necessary in practice.

Last, we gave each participant personalized written feedback, which was provided by professionals in human resources (for the job interviews) and hospitality (for the reception desk). The experts were instructed to give constructive feedback on how the students could improve their performance either at the desk or during the interview, based on the full audio-video recording of the interactions during the first lab session.

3.2.3 Participants

In total, we collected 169 job interviews and reception desk interactions, for a total of 338 interactions. In aggregate, the corpus comprises 3040 minutes of recordings, with 1690 minutes of interviews (average interview duration $\approx$ 10 minutes) and 1350 minutes for desk (average desk duration $\approx$ 8 minutes). To our knowledge, this constitutes one of the largest academic dataset of dyadic interactions collected in an organizational context. 100 students participated in the first lab session, 69 participants completed both lab sessions. The 31 students who did not complete the full training procedure either decided to leave the study before or after the feedback session, or did not reply to emails or SMS. Out of the 100 students who participated in at least one lab session, there were 57 females, and the mean age was 20.6 years (SD = 2.47). Additionally, because the hospitality management school has programs in both French and English, the two languages are present in the corpus. Out of the 169 sessions,

130 were conducted in French.

3.3 Annotations

This is subsection, we present the methodology followed to annotate the data collected in Section 3.2. The annotators were asked to rate the participants they viewed, on a number of social variables as detailed below. The raters, however, were not asked to annotate behavioral cues like hesitations and smiles. Specifically, the raters were asked to answer the question "I consider the person as ..." and had to rate the participant's skill on a seven-point Likert scale. As the same raters annotated various social variables, there is the influence of the "halo effect" . This effect is the phenomenon that causes people to be biased in their judgments by transferring their feelings about one attribute of something to other attributes [160, 44, 42].

3.3.1 Job Interviews

As the objective of this study was to implement a behavioral training procedure for hospitality student, we were interested in the effects of nonverbal behavior in the perception of various social variables such as *Hirability*; *Overall Impression*; *Professional Skills* (Competent, Motivated, Hardworking); *Social Skills* (Enthusiastic, Positive, Sociable); and *Communication Skills* (Communicative, Concise, Persuasive). To observe the initial effect, we showed the first two minutes of the interview video to five raters, who were master's student in psychology. The raters watched the first two minutes of the videos and rated a number of social variables on Likert scale from 1 to 7. The use of thin slices in a common practice in psychology [7] and social computing [132]. The comparison with predictive validity across slices is a research issue for future work.

3.3.2 Reception Desk

We enriched the audio-visual dataset with a number of manually labeled variables. Impressions of performance, skills and personality traits (Big-5) were coded by one group consisting of three independent annotators (Group-A), while attractiveness attributes were rated by a separate group of three independent annotators (Group-B). The choice of two separate groups was motivated by the fact that asking the raters to focus on physical attractiveness could influence performance impressions, the very question under study. The annotators were students, who responded to a call for volunteers and were paid 20 USD per hour for their work.

Annotators in Group-A watched the first two minutes of the complaint segment of all the reception desk videos. These two-minute segments were selected as thin slices, following previous work in psychology [7] and social computing [132, 92]. Annotators rated all the receptionists on a number of impression variables using a seven-point Likert scale. The list of variables (Table 6.1) includes: *Performance* (participant's ability to stay calm, satisfy

customers, be patient and calm, and be resistant to stress); *Overall Impression*; *Professional Skills* (Competent, Motivated, Satisfying); *Social Skills* (Intelligent, Positive, Sociable); and *Communication Skills* (Clear, Persuasive). Several of these variables have been studied in other work-related computational studies [114, 112, 108]; we examine them for the reception desk case. In addition, the annotators were asked to rate the perceived personality traits of all participants. We used the standard Ten Item Personality Inventory (TIPI) consisting of ten items, two per dimension [62]. TIPI is widely used to collect impressions of personality and is easy to administer as it consists of only ten questions. Our goal is not to predict personality, but to use big-5 traits as features to predict job performance, following previous literature.

Annotators in Group-B were asked to rate attractiveness of the participants based on still images. The images were video frames selected based on the following criteria: (1) Full frontal face was visible with no occlusion, and (2) the participant displayed a neutral face (no smiling or any other expression). The attractiveness of each participant was assessed using four variables: *Physical Attractiveness*, *Likeable*, *Dislikeable*, and *Friendly*. The use of both physical and non-physical attractiveness was inspired by its previous use in literature [2, 80, 94, 136, 151]. Annotators were asked to answer the questions: "How attractive do you find this person?" for *Physical Attractiveness*, and similarly for the other attributes. This was rated on a five-point Likert scale which was later rescaled to seven-point scale before analysis. The ratings of six participants were excluded due to technical reasons thus rendering $N = 163$ for all attractiveness related analysis.

3.4 Transcriptions

To investigate the impact of linguistic style employed by participants in each situation, we utilized manually transcribed text from the audio tracks. As the impact of words on perceived job performance has not been previously investigated, we choose manual transcription instead of using an automatic speech recognition (ASR) system to set a gold standard against which future ASR works could be compared with. The transcription was done by a pool of five masters' students from organizational psychology, who were native French speakers and fluent in English, watched all the videos, and transcribed the interaction in the original language. The transcribed documents contained verbal content of both the research assistants' and the participants' speech. In our analysis, we use only the participants' data for two reasons: our focus is on participants behavior, and the research assistants' questions did not vary during the job interview situation.

The average number of transcribed words for an interview (applicant answers only) was 813, with a minimum of 358 and a maximum of 2587 words. For the desk situation, the mean number of words transcribed was 354.1, with a minimum of 140 words and maximum of 1027 words. This difference in the mean number of words is due to the scenario, as the job interview setting needs the participant to speak more, while in the reception desk interaction, the client is unhappy and speaks more.

3.5 Extraction of Behavioral Cues

A number of features were extracted to characterize the nonverbal behavior and verbal content representation of participants. The choice of nonverbal cues and verbal features were guided by existing literature in social psychology [40, 77, 6, 32], hospitality [157] and social computing [92, 114, 108, 28].

3.5.1 Audio-Visual Nonverbal Cues

1. **Acoustic Features** can be divided into two types: speaking activity and prosody. Studies in interpersonal communication indicate that vocal characteristics such as pauses, pitch, loudness are used by listeners to perceive the speakers' intent [144, 149]. The feature vector of this modality is of length 98.

   a **Speaking Activity Features** have been shown in psychology and hospitality literature to be correlated to impression formation in various workplace interactions [40, 22], and have been validated in social computing literature [92, 114]. These features comprises cues like speaking time (total time that an individual speaks), speaking turns (active segments greater than two seconds), pauses (gaps in speech shorter than two seconds), short utterances (speaking segments shorter than two seconds), and silence (gaps in speech greater than two seconds). Research has shown that fluent speech, free of pauses and short utterances is considered more credible than non-fluent speech [46], although brief to moderate pauses have shown to enhance trustworthiness [144]. In contrast, speaking features like silence and disfluencies are associated with anxiety and negative affect [149]. Hence, we selected the above mentioned behavioral cues, and were extracted based on the speaker segmentation provided by a commercial microphone array.

   b **Prosody Features** were extracted from freely available MATLAB code [27, 129]. These features include pitch (voice fundamental frequency), speaking rate (speed at which words are spoken), spectral entropy (measure of irregularity or complexity), energy (voice loudness), voicing rate (number of voiced segments per second), and time derivative of energy (voice loudness modulation). These features have been associated with display of emotions but also play an important part in conveying other social cues like warmth and friendliness. For example, a conversational style of speaking (characterized by lower pitch, slower speaking rate with low to moderate volume) has been shown to be correlated with perception of trustworthiness, warmth, kindness and friendliness, while a dynamic style of speaking (higher pitch, faster speaking with high volume) has been associated with dynamism, dominance and competence [124]. The following statistics were extracted and used as features: mean, standard deviation, minimum, maximum, entropy, median, and quartiles.

2. **Visual Features** are further divided on the basis of overall body motion, head nods, and facial expressions. Also known as *Kinesics*, these cues play an important role

in nonverbal communication and have been shown to influence interviewer's and clients' assessments [73, 99, 157]. Gesturing is associated with passion and contributes to the effectiveness of a message being delivered [23] while head nodding enhances perceptions of empathy, courtesy and trust [157]. Facial expressions are known to be associated with various social behavior including dominance [120, 155], warmth [16] and emotional distance [26]. A number of statistics, including count, mean, median, standard deviation, minimum, maximum, entropy, quartiles, and center of gravity, were computed for use as features. The length of the visual feature vector is 64.

a **Overall Visual Motion** captures the total amount of visual movement displayed during the entire interaction. This feature is computed by a modified version of motion energy images, called Weighted Motion Energy Images (WMEI) [18].

b **Head Nods & Channeling:** Head nods were extracted using a 3D face centered method [31]. In this method, a 3D head tracker to calculate the angular velocities using relative rotation at each instant with respect to the head pose at some earlier instance. This method provided a per frame output, with nodding indicated with 1 and no nods indicated by -1. We define visual back-channeling (visual BC) as an event when a person nodded while the other was speaking. This cue was obtained by synchronizing the speaking activity of both participant and the interactor with head nod activity. Another cue we extracted is nodding while speaking. We define these as the occurrence of protagonists nodding their head while speaking.

c **Eye Gaze:** Gaze features were extracted by doctoral student Remy Siegfried as part of a collaboration described in [111] paper. Gaze features were extracted using an off-the-shelf method to compute participants' gaze (i.e. line of sight in 3D space) from a Kinect v2 sensor [52]. Taking advantage of the context (i.e. a human-human interaction), we improved its outputs using a subject adaptation method that computes the gaze estimation bias based on speaking turns [148]. An overview of this method is presented in Figure 3.4 and follows the intuition that people are more likely to look at the person who is speaking during a conversation. The gaze errors corresponding to the frames where the other person is speaking and where the estimated gaze and head pose make a glance at the speaker possible are stored. Then, the bias is estimated using the Least Median Square (LMedS) estimator, so that remaining outliers frames (i.e. frames where the subject is not looking at the other person despite the fact that he/she is talking) are removed. Finally, the bias can be subtracted from the gaze signal, which allows us to compute the Visual Focus of Attention (VFOA) with an higher accuracy without the need to for manual calibration or annotation. This method allowed to reach a mean error of 7.64° on annotated subsets of the interview and desk datasets.

Using this data, we extract the various eye gaze cues, defined as number and duration of events when participant was gazing at the protagonist, from both client and participant.These include gaze while speaking (*GWS*; number and duration of looking at the protagonist and speaking), gazing while listening (*GWL*; number and duration of looking at the protagonist and not speaking), visual dominance ratio (*VDR* defined as

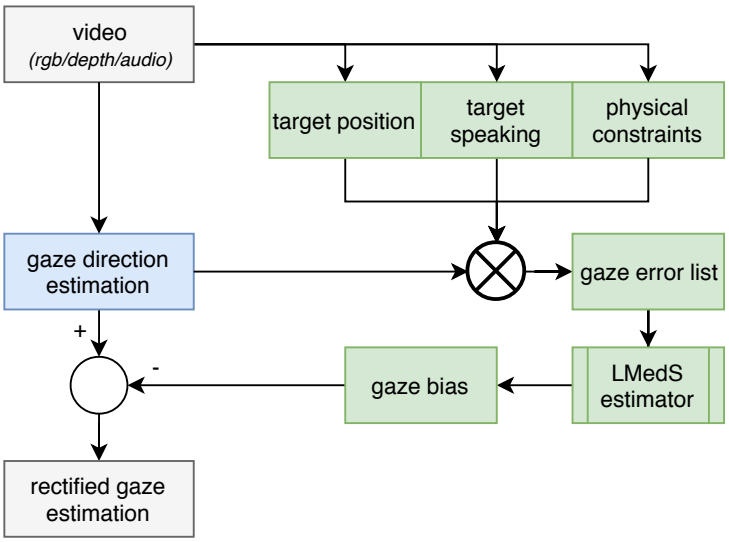


Figure 3.4 – Gaze bias correction method using speaking turns.

ratio of GWS and GWL [43]), and mutual gaze.

d **Facial Expression Features** were extracted using the Emotion API of Microsoft Azure cognitive services [154, 153]. Various cloud-based services [146] are available and has been previously used in the literature to study diverse social, political, and information interaction issues including cyber-bullying [150] and public health images [56]. In this work, we use the Microsoft Azure Emotion API [34] to extract emotions from facial expressions. As a first step, video frames were extracted from video clips at 5 frames per second (fps). Then, these images were input sequentially to the API. The output was confidence values across 8 facial expressions of emotion (happiness, sadness, surprise, anger, fear, contempt, disgust and neutral) normalized to 1. If a face was not found, the API returned 0 for all values and were filtered before processing. Various statistics were computed from this 8-dimensional vector and used as features.

3.5.2 Verbal Cues

3. **Linguistic Features:** Peoples' choice of words while speaking and writing reveal aspects of a person's identity [32] while also providing cues to their thought processes, emotional states, intentions, and motivations [127, 159]. Although the impact of linguistic style on perceived hirability and soft skills have investigated in the literature [112, 28, 106], little is known about the role of linguistic content on perceived performance and soft skills.

   a **Lexical Features** were extracted using Linguistic Inquiry and Word Count (LIWC) [127], a software widely used in social psychology [32] and social computing [20, 143]. LIWC looks up each word in the interview transcript to the dictionary and then maps them to one of 71 categories (e.g proper pronouns, adjectives, verbs etc) and increments the appropriate word category. It must be noted that LIWC can assign words to more

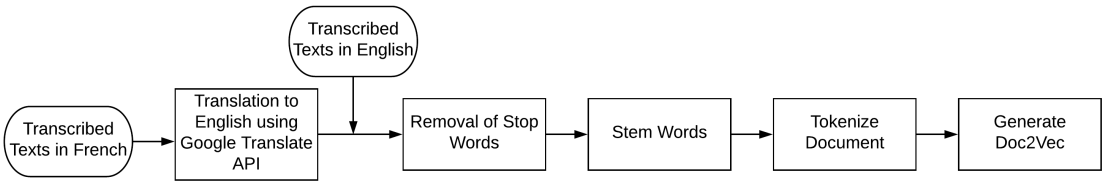


Figure 3.5 – Flow chart showing the steps followed for extracting Doc2Vec features from transcribed text of both the situations.

than one category at a time. After processing a document, LIWC divides the count of categories by the total number of words in the document hence normalizing them. LIWC is designed to process raw text, thus no pre-processed of transcripts was required. The total length of this features is 70.

c **Doc2Vec** is a python implementation [135] of the paragraph vector method [89]. Doc2Vec generates a fixed length vector for numerical representation of text data of varying size, such as sentences, paragraphs, or documents. The steps taken to extract Doc2Vec features is summarized in Figure 3.5. Our data corpus consists of two languages (French and English) so as a first step we translated all the French text into English. This translation was done using the Google Translate API, a statistical translation system in which language models are trained on billions of words of equivalent text in different languages. This API was found to be the most accurate in three of four tests consisting of 20 machine translation systems [118]. This is a necessary step as otherwise the word embedding trained would be in two separate spaces making any comparisons difficult. We then remove stop words from the text data [161]. In the next step, the text is converted into lower case, stemmed and tokenized using the NLTK package [21] in Python. We then generate document vectors by training a model for word embedding using the Gensim package [135]. The model was generated by selecting a constant learning rate for 10 epochs with 100 iterations and a vector of length 100. The choice of these parameters was guided by the small corpus size.

3.6 Conclusion

In this chapter, we described a number of challenges associated with implementing a behavioral training procedure for hospitality students in order to improve the first impressions that others form about them. The framework consisted of two scenarios that were relevant for their future careers, namely job interviews and reception desk situations. During this implementation, we collected the UBIMPRESSED data corpus. The data corpus consisted of 169 dyadic interactions each in two hospitality settings; job interviews and reception desk and was recorded with multiple modalities. Nonverbal cues were automatically extracted and their relationship with various perceived social variables analyzed. The interactions were manually coded to obtain impressions of various social variables. The interactions were also

manually transcribed with an aim of understanding connections between choice of words and formation of first impressions in and across two workplace situations.

4 First Impressions in Employment Interviews

First impressions are key in the context of organizations, such as during employment interviews, but also while working on jobs requiring strong communication skills such as sales, marketing, or hospitality. First impressions can be defined as the mental image one forms about something or someone after a first encounter. In professional spheres, these initial judgments can influence critical outcomes, such as being hired or promoted. Psychology researchers have shown that people can make accurate inferences about others, even if the amount of information is very limited [5]. Nonverbal behavior has been established as a major channel through which information is communicated; it constitutes a strong basis on which first impressions are formed [87]. In this chapter, we describe our analysis of the job interview part of our data corpus (Section 3.2).

Job interviews are ubiquitous and the impact of nonverbal and verbal behavior (NVB) on job interview outcomes has been studied in psychology [13, 40, 47, 137] and social computing [114, 108, 112, 28]. Organizational psychologists have studied job interviews for decades, aiming at understanding the relationships among personality, behavior (both verbal and nonverbal), interview ratings/outcomes, and job performance [40, 47, 137]. Until recently, these studies have been conducted based on manual annotations of behavior, which is labor-intensive and makes it difficult to scale. In the last decade, the advent of inexpensive sensors combined with improved perceptual techniques have enabled the possibility to automatically analyze human face-to-face interactions [58, 130]. In the context of job interviews, recent studies have established the feasibility of automatically inferring interview ratings [114] and other related constructs (e.g. engagement, friendliness, or excitement) [112] up to a certain level. Other researchers [72] have extended these works by developing a social coaching system to enable potential job-seekers to train their behavior in order to convey a more positive first impression to recruiters. To this end, they provided feedback to college undergraduate students about their automatically sensed nonverbal behavior, including head gestures, smiles, and prosody. The subjects that obtained feedback via the coaching system showed improved performance during interviews.

The contributions of this chapter are the following. We analyzed the relationship between

automatically extracted nonverbal cues and various perceived social variables in a correlation analysis and a prediction task, and extended the results found in [114]. We found gender differences that confirmed previous findings in psychology [90]. We then extracted verbal cues from the job interview transcripts and analyze the relationship between the verbal cues and impressions of professional, communication and social skills through correlation analysis. We then defined a regression task to infer the overall impression scores and other skill variables by utilizing the extracted verbal and nonverbal cues. To understand the underlying structure of the skill impressions, we then performed a Principal Component Analysis (PCA). on the annotated ratings of these social variables. We found that the projection of the skill impressions onto the first principal component, could be inferred with $R^2 = 0.41$.

This chapter is organized as follows. Section 4.1 briefly describes the interview part of our data corpus and the annotations. Section 4.2 describes the various features extracted from this set of videos. In Section 4.3, we describe our motivations and findings of a Principal Component Analysis of the annotated skills. Section 4.4 and Section 4.5 outline the various correlation and inference experiments conducted. The material of this chapter was originally published in [108, 106, 111].

4.1 Dataset

In this chapter, we use the job interview part of the data corpus described in Section 3.2. Here we briefly outline the scenario, data collection method, annotations and nonverbal features extracted to keep the chapter self-contained.

4.1.1 Data Corpus

This dataset consists of 169 job interview interactions previously collected by us a part of the UBIMPRESSED project. The interviews were recorded synchronously using (1) two Kinect cameras (one for each protagonist) at 30 frames per second; (2) a microphone array placed at the center of the table recording audio at 48kHz. The annotations for interview videos was done by 5 raters. The raters watched the first two minutes of the videos and annotated the variables on a seven-point Likert scale. Use of two-minute segments, also known as thin slices, follows existing literature [7, 132]. Details of the data collection can be found in Chapter 3).

The agreement between the raters was assessed using Intraclass Correlation Coefficient (ICC), a statistics that describes how strongly units in the same group resemble each other [147]. ICC is a standard measure of inter-rater reliability widely used in psychology and social computing [49, 114] As all the raters annotated each video and because we used a sample rather than a population of raters, we used a two-way mixed, consistency, average-measures $ICC(2,k)$. The agreement between raters for all the social variables was greater than 0.5 indicating moderate reliability. Table 4.1 summarizes the annotated social variables and presents their respective descriptive statistics.

Table 4.1 – Intra class correlation ($ICC(2,k)$) and descriptive statistics for perceived social variables in the job interview data corpus. The agreement between raters for all the social variables was greater than 0.5 indicating moderate reliability. Furthemore, the mean ratings of all social variables is greater then 4 indicting that the participants were overall percieved positively.

Social Variable	**ICC(2,k)**	**mean**	**std**	**skew**
Professional Skills				
Motivated (motiv)	0.52	5.89	0.60	-1.03
Competent (compe)	0.56	6.01	0.54	-1.00
Hardworking (hardw)	0.54	6.06	0.55	-1.07
Social Skills				
Sociable (socia)	0.57	5.67	0.65	-0.39
Enthusiastic (enthu)	0.68	5.52	0.87	-0.64
Positive (posit)	0.60	5.70	0.72	-0.46
Communication Skills				
Communicative (commu)	0.60	5.82	0.71	-0.79
Concise (consi)	0.55	5.84	0.65	-0.72
Persuasive (persu)	0.69	5.57	0.87	-0.76
Overall Impression				
OvImpress	0.73	5.58	0.94	-0.76

As a first step, we analyzed the pairwise correlations (using Pearson's correlation) between the social variables. These are presented in Table 4.2. All variables annotated were observed to be significantly correlated with each other with correlation coefficients above 0.6 for all cases. Correlations between some variables like competent and hardworking (r = 0.91), sociable and enthusiastic (r = 0.91), enthusiastic and positive (r = 0.96) were very high, indicating that they are essentially the same.

Table 4.2 – Correlation matrix for perceived social variables in job interviews (N=169) . In all cases, correlation is significant (p<0.01).

	2	3	4	5	6	7	8	9	10
1. ovImpression	0.87	0.85	0.83	0.79	0.85	0.87	0.83	0.81	0.88
2. motivated		0.85	0.84	0.76	0.83	0.84	0.79	0.65	0.74
3. competent			0.92	0.62	0.71	0.73	0.67	0.68	0.74
4. hardworking				0.60	0.69	0.72	0.66	0.64	0.69
5. sociable					0.91	0.89	0.83	0.64	0.76
6. enthusiastic						0.96	0.89	0.69	0.81
7. positive							0.90	0.73	0.81
8. communicative								0.75	0.85
9. concise									0.88
10. persuasive									

4.1.2 Transcripts

To analyze the linguistic content, we manually transcribed all the interviews in the data corpus using a pool of five master's students in organizational psychology, who were native French speakers and fluent in English. Each question by the interviewer and answer by the applicant was transcribed, but only the applicant answers were utilized for linguistic analysis. The average number of transcribed words for an interview (applicant answers only) was 813, with a minimum of 358 and a maximum of 2587 words. In aggregate, the corpus comprised of 1690 minutes of interviews (mean duration: 10 minutes).

4.2 Verbal and Nonverbal Features

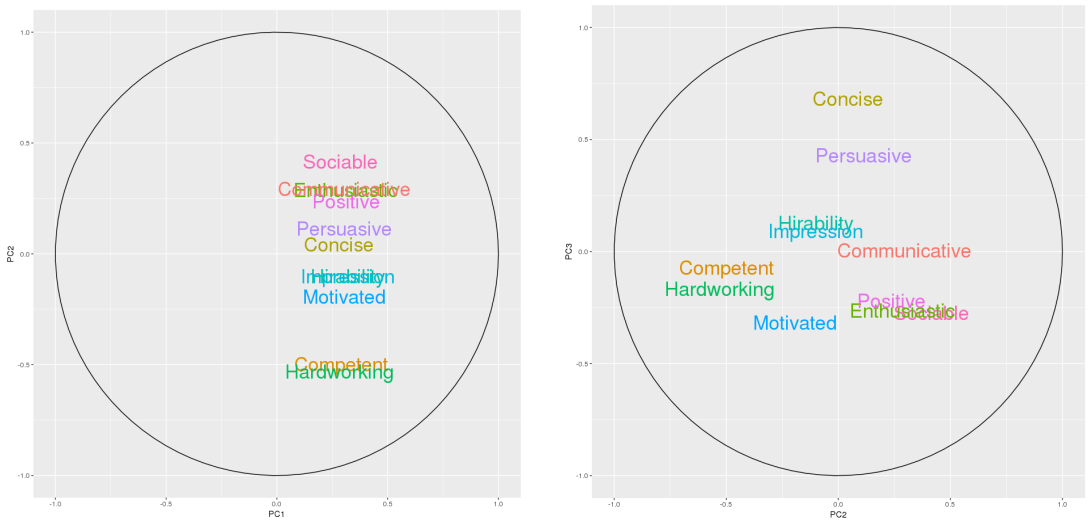
In this section, we outline the various nonverbal and verbal features used in this chapter. A more detailed description of the features and their use in literature can be found in Section 3.5.

Nonverbal Feature: We extracted various nonverbal cues from both visual and audio modality for behavioral representation of the dyadic interaction. These nonverbal cues include prosodic cues (pitch, energy, voice loudness modulation, spectral entropy), speaking activity features (speaking time, speaking turns, pauses, short utterances), visual motion (WMEI [18]) and head nods [31]. A detailed description of the extracted features can be found in Section 3.5. The choice of nonverbal cues extracted was based literature in social psychology [39, 77] and social computing [114, 115]. The nonverbal cues were extracted for the full interview and for both applicant and interviewer.

Verbal Features: To understand the connection between choice of words and first impressions in job interviews, we processed the interview transcripts to extract lexical features with the Linguistic Inquiry and Word Count (LIWC) [127], a software package widely used in social psychology [32] and social computing [20, 143] to extract verbal content.

4.3 Principal Component Analysis of Skills

As shown in [108], the manually annotated impression variables are correlated. Pairwise Pearson correlations was found to be in the range $r \in [0.60, 0.96]$ (median = 0.81). This suggests that a lower dimensionality representation of impressions could be found through principal component analysis (PCA) of the annotated variables. As a first step towards PCA, the annotated values were first pre-processed so that each variable had zero mean and unity variance. Then, PCA was conducted on the skill variables using the inbuilt *prcomp* function in R. The first three principal components (PC) are visualized in Figure 4.1, which displays the original variables projected onto the coordinate system. We observe that these three PCs account for 94.3% of the variance. Further, we observe that the first component, accounting for 82.6% of the variance, essentially corresponds to overall positive and negative impressions. Note that all the variables in Table 4.1 are positively phrased. We also observe that *Communicative* and



(a) (b)

Figure 4.1 – Clustering of perceived skills after principal component analysis (PCA). The first three principal components, accounting for 94.3% of the variance, are displayed here by projecting the original axes onto the PCA space ($N = 169$).

Enthusiastic overlap, and *Hirability* and *Overall Impressions* overlap. The second principal component, accounting for 6.1% of the variance, seemed to distinguish professional skills (Motivated, Competent) from communication (Communicative, Concise) and social (Positive, Sociable) skills (Figure 4.1b). The third component accounted for 5.6% of the variance. We can thus see that this lower dimensional representation of the impressions is appealing. The projection onto the first PCs were used as the labels in an inference task (Section 4.5.4).

4.4 Correlation Analysis

In the section, we use nonverbal and verbal features extracted from both applicant and interviewer in our data. Using perceived soft skills and hirability as predictors, we investigate the linear relationship between behavioral cues and predictors. The results of correlation analysis for the social variables are presented in Table 6.7.

4.4.1 Overall Impression & Nonverbal Cues

Applicant cues

A number of applicant features were found to be significantly correlated to impressions of all social variables of interest. Specifically, participants who spoke often, longer, louder, with greater modulation of loudness and pitch obtained higher score in overall impression, professional, social and communicative skills. This results are in accordance with existing

Table 4.3 – Selected Pearson's correlation coefficient for various social variables.* $p < 0.01$, † $p < 0.05$

Nonverbal Cues	Professional Skills			Social Skills			Communication Skills			Overall
	motiv	compe	hardw	socia	enthu	posit	commu	conci	persu	ovImpress
Applicant Speaking Activity										
Avg Turn duration	0.38*	0.27*	0.27*	0.39*	0.41*	0.39*	0.38*	0.20†	0.32*	0.36*
Num Silent Events	−0.39*	−0.27*	−0.26*	−0.43*	−0.48*	−0.47*	−0.41*	−0.32*	−0.35*	−0.42*
Applicant Pitch										
Std Deviation	−0.19†	−0.15	−0.22*	−0.21†	−0.28*	−0.29*	−0.22*	−0.18†	−0.16	−0.24*
Lower quartile	0.21*	0.22*	0.28*		0.26*	0.27*	0.19†	0.17†		0.25*
Applicant Spectral Entropy										
Average	−0.14	−0.16†	−0.20†		−0.16†	−0.18†		−0.23*		−0.23*
Std Deviation	0.26*	0.29*	0.34*	0.18†	0.30*	0.31*	0.22*	0.21*	0.17†	0.28*
Applicant Energy										
Average	0.37*	0.24*	0.24*	0.33*	0.36*	0.36*	0.30*	0.19†	0.28*	0.31*
Std Deviation	0.39*	0.27*	0.27*	0.37*	0.40*	0.40*	0.35*	0.22*	0.31*	0.34*
Lower quartile	0.34*	0.27*	0.23*	0.32*	0.33*	0.31*	0.30*	0.23*	0.33*	0.30*
Maximum	0.41*	0.30*	0.31*	0.40*	0.44*	0.43*	0.40*	0.22*	0.33*	0.35*
Applicant Change in Energy										
Maximum	0.41*	0.31*	0.32*	0.44*	0.46*	0.44*	0.40*	0.22*	0.33*	0.35*
Minimum	−0.40*	−0.27*	−0.30*	−0.42*	−0.44*	−0.43*	−0.40*	−0.20†	−0.31*	−0.35*
Applicant WMEI										
Maximum	0.26*	0.26*	0.24*	0.15	0.17†	0.15	0.18†			0.16
Interviewer Pitch										
Average				0.14	0.15	0.15				0.15
Std Deviation	−0.15			−0.19†	−0.19†	−0.21†	−0.14	−0.19†	−0.19†	−0.22*
Interviewer Spectral Entropy										
Std Deviation	0.17†	0.15		0.23*	0.24*	0.24*	0.17†	0.19†	0.20†	0.24*
Minimum	−0.21*	−0.16		−0.23*	−0.25*	−0.26*	−0.14	−0.21*	−0.18†	−0.26*
Interviewer Energy										
Std Deviation	0.28*	0.19†	0.19†	0.24*	0.25*	0.25*	0.25*		0.16	0.19†
Maximum	0.28*	0.21*	0.23*	0.25*	0.29*	0.27*	0.26*		0.19†	0.21*
Interviewer Change in Energy										
Maximum	0.28*	0.24*	0.24*	0.25*	0.29*	0.28*	0.26*	0.14	0.20†	0.22*
Minimum	−0.27*	−0.19†	−0.21*	−0.27*	−0.29*	−0.27*	−0.25*		−0.19†	−0.21*
Interviewer WMEI										
Average	0.23*	0.21*	0.19†		0.15	0.15	0.15	0.14		0.16

literature [39, 45, 114]. Similarly, applicants who spoke animatedly with more gestures and motion were more favorably viewed than participants who spoke with less gestures. This corroborates the results in [114], which showed that applicants displaying more visual head motion (WMEI) received better hireability scores.

The prosodic cues related to energy (max, std, mean and lower quartile), time derivative of energy (maximum, std, upper quartile), and pitch (lower quartile), were also found to be significantly positively correlated with the overall impression scores. While time derivative of energy (min), spectral entropy (min, lower quartile and mean),and pitch (std, entropy), which were found to be negatively, significantly correlated with to all the social variables of interest. This implies that participants who spoke louder, with greater modulation of loudness and pitch were more favorably viewed than participants who spoke with less voice modulation. This results too have been confirmed in previous literature [39, 45]. while participants who spoke monotonously with less modulation or change in pitch were rated lower. when extracted from both the full interview and most thin slices voiced rate (mean, std, median, and upper quartile), professional and social skills. Specifically, participants who spoke longer with greater energy with less silence obtained better impression ratings. This is in accordance with previous findings []. An applicant's maximum visual motion was found to be moderately correlated with their overall impression, profession, social skills overall impressions or any other social variables of interest. This is in accordance with the results presented in [114] which showed that applicants who displayed more visual head motion received better hireability scores. We

Table 4.4 – Range of Pearsons correlation between eye gaze, facial expressions and social variables in the interview ($N = 161$) dataset. $^{***}p < 0.001$;$^{**}p < 0.01$;$^{*}p < 0.05$.

	Professional	Social	Communication	Hirability
Applicant				
Std Duration (GWS)	.15	[.15, .18*]	[.18, .22]**	.17*
Max Duration (GWS)	[.16, .17]*	[.16, .19*]	[.18, .22]**	.18*
Num of GWL	[−.18*, −.11]	[−.25, −.28]***	[−.30, −.23]**	−.22**
Mean Duration (GWL)	[.15, .23**]	[.12, .17*]	[.21, .23]**	.16*
Std Duration (GWL)	[.13, .16]	[.17, .20]*	.22**	.20*
VDR	[.22, .28]***	.27***	[.17, .25]**	.28***
Facial Expression				
Mean Neutral	[−.32, −.22]**	[−.43, −.40]***	[−.24, −.22]**	−.30***
Median Neutral	[−.27, −.20]**	[−.33, −.30]***	−.18*	−.24**
Std Neutral	[.15, .28***]	.43***	[.19, .24]*	.28***
Var Neutral	.12, .27***	.42***	[.18, .23]*	.26***
Mean Happy	.31***	[.41, .42]***	[.22, .26]**	.30***
Median Happy	.25**	[.27, .28]***	.19*	.22**
Std Happy	.25**	.41***	[.16, .22]*	.25**
Var Happy	.25**	[.40, .41]***	[.16, .22]*	.25**

believe this can be explained by the fact that applicants for hospitality positions are expected to be less exuberant and more formal than other job interviews.

We observe low to moderate correlations between eye gaze, facial expressions and perceived soft skills (Table 4.4). Specifically, we observe that applicants who had greater VDR (i.e amount of visual dominance) and presented a happy expression were perceived to be more hirable than applicants who presented a neutral expression and displayed lower VDR (i.e looked less at the interviewer while speaking). Our results concur with literature, which report positive correlation between eye gaze and hirabilty [48, 77, 3], and happy and hirability [77, 29].

Interviewer cues

Interviewer's pitch (std), spectral entropy (min, lower quartile) and time derivative of energy (minimum) were observed to be negatively associated with all social variables suggesting that the interviewer had a more monotonous tone of voice in presence of job applicants who were rated higher. These results is in line with those reported in [39, 114]. Interviewer's overall visual motion (mean) was positively associated with overall impression and all social variables of interest. This suggests that interviewers gestured more in the presence of an applicant who scored higher than applicants with lower scores, thus again validating the findings in [114]. Interestingly, the interviewer gaze and expression cues were not correlated to perceived hirability and performance.

Table 4.5 – Correlation between linguistic cues and impressions in job interviews (N = 169) with significance values ** $p < 0.001$; * $p < 0.05$. Others not significant.

SocVar	Personal Pronoun	1st pers singular	3rd pers singular	3rd pers plural	Negation	Non fluency	Question
motiv	**-0.17***	-0.12	**-0.21****	0.12	-0.10	-0.15	-0.09
compe	-0.15	-0.10	**-0.25****	**0.22****	**-0.21****	**-0.25****	-0.08
hardw	-0.14	-0.07	**-0.27****	**0.24****	-0.11	**-0.17***	-0.08
socia	-0.12	-0.12	-0.14	0.11	-0.09	-0.11	-0.15
enthu	-0.13	**-0.16***	**-0.16***	**0.17***	-0.13	-0.12	**-0.17***
posit	-0.14	-0.13	**-0.19***	**0.17***	-0.14	-0.12	-0.13
commu	-0.13	-0.13	-0.11	0.11	-0.13	-0.08	**-0.18***
conci	**-0.22****	**-0.16***	**-0.15***	0.12	**-0.17***	**-0.25****	**-0.18***
persu	**-0.18***	**-0.20***	-0.13	0.14	**-0.16***	**-0.20***	**-0.19***
ovImp	**-0.18***	**-0.17***	**-0.17***	**0.17***	**-0.17***	**-0.17***	**-0.20****

4.4.2 Overall Impression & Verbal Behavior

To understand the connection between words used and formation of first impressions, we conduct a pairwise correlation (using Pearson's correlation) between all the impression variables and verbal cues extracted. For this analysis, the mean rating of all variables by the five raters and the features extracted from LIWC were utilized. Table 4.5 shows that a few LIWC features significantly correlate with overall impression scores. We observe that there is a low correlation between verbal content and *Overall Impression* (ovImp). The use of personal pronouns (*pproun*), 1st person singular (*i, me*), 3rd person singular (*she, him*) are negatively correlated to overall impressions while use of 3rd person plural (*they, their*) are positively correlated. These observations are somewhat in line with those of reported by authors in [112]. Other weak effects observed are that participants who used negation (*negate*), non fluency (*hmm, er*) and asked questions (*QMark*) had lower overall impression scores.

4.5 Regression Analysis

We divide our experiments into three sections. In Section 4.5.1 and Section 4.5.3, we evaluate a computational framework for the automatic inference of first impressions from employment interviews using nonverbal and verbal cues respectively. In Section 4.5.2, we present results comparing ratings and behavior between the first and second interviews.

Several regression techniques (Ridge, random forest (RF), ordinary least squares (OLS)) were evaluated. For these tasks, we used leave-one-interview-out cross validation and 10-fold inner cross validation. As we used the leave-one-interview-out cross validation, it is possible that one participant can be in both the training and test sets although in different interviews. For evaluation measure, we utilized the coefficient of determination R^2, which accounts for the amount of total variance explained by the model under analysis. This metric is often used in both psychology and social computing to evaluate regression tasks.

Table 4.6 – Regression results for each language and gender using regression methods; pVal with random forest (pVal-RF) and pVal with Ridge (pVal-Ridge) $^*p < 0.01$, $\dagger p < 0.05$.

Nonverbal Cues	All ($N = 169$)		English ($N = 39$)		French ($N = 130$)		Female ($N = 96$)		Male ($N = 73$)	
	Method	R^2	Method	R^2	Method	R^2	Method	R^2	Method	R^2
Overall Impression	pVal-RF	0.32^*	pVal-Ridge	$0.14\dagger$	pVal-RF	0.32^*	pVal-Ridge	0.06	pVal-RF	0.44^*
Communication	pVal-RF	0.25^*	pVal-Ridge	$0.07\dagger$	pVal-Ridge	$0.16\dagger$	pVal-Ridge	−0.07	pVal-Ridge	0.45^*
Persuasive	pVal-RF	$0.20\dagger$	pVal-Ridge	$0.09\dagger$	pVal-RF	$0.20\dagger$	pVal-RF	0.05	pVal-RF	$0.28\dagger$
Concise	pVal-RF	$0.14\dagger$	pVal-Ridge	$0.38\dagger$	pVal-Ridge	−0.06	pVal-Ridge	−0.13	Pval-RF	$0.13\dagger$
Enthusiastic	pVal-RF	0.34^*	pVal-RF	$0.12\dagger$	pVal-RF	0.31^*	pVal-Ridge	0.07	pVal-Ridge	0.46^*
Positive	pVal-RF	0.30^*	pVal-RF	$0.06\dagger$	pVal-RF	$0.27\dagger$	pVal-Ridge	0.12	pVal-Ridge	0.44^*
Social	pVal-RF	0.19^*	pVal-RF	$0.14\dagger$	pVal-RF	0.15	pVal-Ridge	0.06	pVal-Ridge	0.44^*
Competence	pVal-RF	0.18	pVal-Ridge	$0.15\dagger$	pVal-RF	$0.22\dagger$	pVal-Ridge	−0.10	pVal-Ridge	0.38^*
Hardworking	pVal-RF	0.15	pVal-Ridge	−0.24	pVal-RF	0.12	pVal-Ridge	−0.18	pVal-RF	0.44^*
Motivated	pVal-RF	0.29^*	pVal-RF	$0.26\dagger$	pVal-RF	$0.27\dagger$	pVal-Ridge	0.04	pVal-RF	0.44^*

4.5.1 Inferring First Impressions from Nonverbal Cues

The results of the regression task are summarized in Table 4.6. Results from utilizing all the data indicate that all social variables annotated were predictable to some degree from nonverbal behavior. Random forest with pVal dimensionality reduction (pVal-RF) was found to perform best with this set of features. Overall impression and enthusiastic had the highest R^2, 0.34 and 0.32 respectively. This implies nonverbal behavior is predictive of overall first impression as shown in existing literature [87] and corroborates the results found in [114]. We observe certain variables like hardworking ($R^2 = 0.15$) and competence ($R^2 = 0.18$) to be harder to predict.

Comparing to recent work, in [114], the authors reported $R^2 = 0.36$ for hireability, a measure we have not used. They also reported $R^2 = 0.10$ for persuasive and no predictability for communicative. In another work, [112] reported results on a different set of social variables using correlation coefficient r as their evaluation metric. 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 $R^2 = 0.49$. We compare results of socials constructs which are similar in meaning to the ones we have utilized: excited ($R^2 = 0.65$) vs enthusiastic ($R^2 = 0.34$), friendly ($R^2 = 0.63$) vs social ($R^2 = 0.19$), focused ($R^2 = 0.31$) vs motivated ($R^2 = 0.29$). There is no direct way of assessing where the performance difference come from, as the data set used by Naim et al. is not publicly available to our knowledge.

Results of the inference task using the interview data is tabulated in Table 4.8. We observe that eye gaze individually has low inference performance with best of $R^2 = 0.1$ for *Hirability*. Similarly, the performance of facial expressions is low for *Positive* ($R^2 = 0.13$) and *Social* ($R^2 = 0.17$) and poor for other variables. Combining these two cues, we observe an improved inference performance for some variables. Specifically, for *Positive* ($R^2 = 0.21$) followed by *Social* ($R^2 = 0.18$), *Hirability* ($R^2 = 0.13$), and *Persuasive* ($R^2 = 0.11$).

To further understand the impact of eye gaze and facial expressions, we fuse these features with other visual and auditory features. When fused with other visual cues, we observe further improvement in inference performance for *Positive* ($R^2 = 0.22$) followed by *Hirability*

Table 4.7 – Correlation coefficients between selected nonverbal cues and overall impression for sub-sets separated based on gender. $^{*}p < 0.01$; $\dagger p < 0.05$

	Overall impression	
NVB cues	**Female** ($N = 96$)	**Male** ($N = 73$)
App. # of turns	-0.36^{*}	-0.42^{*}
App. speaking time	-	$0.23\dagger$
App. speaking ratio	-	0.35^{*}
App. turn duration stats	$0.25\dagger$	$[0.37, 0.48]^{*}$
App. speech energy stats	$[0.17, 0.26]\dagger$	$[0.42, 0.50]^{*}$
Silence stats	-0.23^{*}	$[-0.47, -0.67]^{*}$

Table 4.8 – Regression analysis using eye gaze and other visual features for job interviews ($N = 161$).

	Clear	Persuasive	Positive	Social	Competent	Motivated	Hirability
Gaze	0.06	0.08	0.06	0.02	0.00	0.04	0.10
Facial Expressions	0.0	0.01	0.13	0.17	0.0	0.04	0.04
Gaze + Expressions	0.06	0.11	0.21	0.18	0.02	0.10	0.13
All Visual	0.04	0.15	0.22	0.22	0.06	0.11	0.14
All Features	0.19	0.30	0.39	0.32	0.20	0.30	0.34
Baseline	0.14	0.20	0.30	0.19	0.18	0.29	0.29

($R^2 = 0.14$), *Social* ($R^2 = 0.22$), and *Persuasive* ($R^2 = 0.15$). Similarly, fusing gaze and expression features with all the previously extracted features (visual + auditory) shows a moderate improvement in inference performance. Particularly, the improvement was highest for *Social* ($R^2 = 0.32$), followed by *Persuasive* ($R^2 = 0.30$), *Positive* ($R^2 = 0.39$), and *Hirability* ($R^2 = 0.32$).

The results of inference task using eye gaze and facial expressions are in agreement with those in psychology and computing literature. Parsons et al. [122] and Forbes et al. [48] showed that high levels of eye contact had a positive effect on interview outcomes. Chen et al. [28], using an off-the-shelf emotion detection toolkit extracted eye gazes, facial expression, and head postures. The authors, using a visual doc2vec representation for features, reported a correlation of $r = 0.36$. For comparison, we convert r to R^2 (by squaring) obtain $R^2 = 0.13$ and is similar to the results in this work. Overall, we see that gaze and facial expressions have a low-moderate effect on inference performance individually and contribute to improved inference when fused with other features.

We observe that language has an effect on the predictive power of impressions scores. A larger variance for overall impression ($R^2 = 0.32$) could be explained for participants using French as the language of interview, while for participants using English only 14% of variance could be explained. We hypothesize that this could be due to the fact that raters were not native English speakers. The same hypothesis could explain why concise had higher R^2 for English than French. However, the small size of the English dataset prevents drawing firm conclusions on the effect of language and will be investigated.

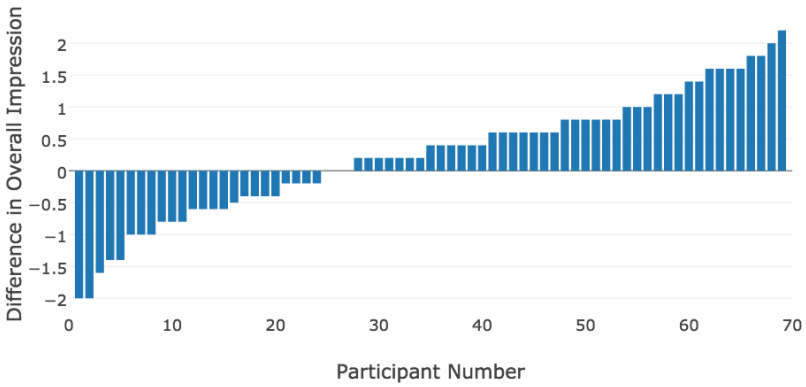


Figure 4.2 – Difference in Overall Impression score between two lab sessions for all participants.

Although no significant difference could be observed between males and females in terms of the values of annotations, we observe that the interviews including a male participant were predicted with higher accuracy ($R^2 \in [0.13, 0.46]$) than the ones featuring females ($R^2 \leq 0.12$). In order to understand these differences, we analyzed the correlations between nonverbal cues and the annotated variables for data subsets separated based on gender. Table 4.7 displays a summary of the largest differences in correlation values for the variable of overall impression; due to space constraints, we did not include the correlations for other variables, but similar trends were found. We observe that the improvement in prediction accuracy can be explained by the overall higher correlations observed for male interviews. Furthermore, striking gender differences can be observed in terms of correlation values, such as speaking time, statistics of turn duration, speech energy, and silence; these behaviors are part of what psychologists refer to as powerful speech. On the one hand, if men use these cues they are perceived as powerful, which is in line with gender stereotypes for men, as well as with persuasiveness [90]. On the other hand, stereotpyically women are less expected to show powerful speech, which might explain the lower correlations found for women.

4.5.2 Changes Across Sessions

To determine if there was a difference in annotated ratings across sessions, we selected only those participants who had completed both lab session 1 and 2 ($N = 69$) from the full dataset. We then conducted a paired t-test on this split dataset, which rejected the null hypothesis for overall impression ($p < 0.05$). Improvement of scores for other social variables were also significant ($p < 0.05$).

Similarly, we conducted a similar experiment to determine if there was a difference in non-verbal behavior across sessions. We observed that there was a significant ($p < 0.01$) change in applicant maximum turn duration and speaking time. Correlation between the change in overall impression scores and change in speaking activity was statistically significant.

Table 4.9 summarizes the descriptive statistics of social variables and nonverbal behaviors which changed significantly between lab session 1 and 2. The difference in overall impression between the two lab sessions for each participant can be visualized in Figure 4.2. We observe that while the majority of the participants had an improvement in their ratings, 34% of participants had a decrease in scores between in the lab session 2 and lab session 1. There was no changes in scores for 4 participants. Although we observe that students overall improved their interview performance at the second laboratory session, no conclusion can be drawn about the factors that were favorable to the student's behavioral improvement. For instance, the source of the improvement in interview performance could be due to the feedback they were given, but also to the fact that they participated to the role-play a second time, in which their level of confidence was higher. In future work, we plan to understand the factors that encourage behavioral improvement: What is the best way to display feedback? What behaviors should feedback focus on? Who are the students that benefit from feedback, e.g. in terms of personality, gender, age? We believe that this work constitutes a first step towards addressing these research questions.

4.5.3 Inferring First Impressions from Verbal Cues

Table 4.10 shows the performance of the RF models for the inference of impressions. We observe that performance of verbal cues in inference of social variables is low ($R^2 \in [0.03, 0.17]$), with the best performance achieved for *Competent* (compe). This results are slightly better then the values reported in [113], and are similar to those reported in other settings like inference of leadership [143], mood [142] and personality [20]. In the job interview setting, the work in [28] used a corpus consisting of 36 participants, extracted verbal cues used LIWC and a Doc2Vec method. Using Pearson's correlation r as their evaluation measure, they reported r = 0.39 with Ridge regression. For comparison, converting r to R^2 (by computing the square of correlation coefficient r) indicates R^2 = 0.15 which is higher than our results.

In comparison, the model trained on nonverbal cues performs better ($R^2 \in [0.21, 0.35]$) for the same dataset. Combining the nonverbal and verbal cues leads to an marginal increase in performance of inference for some social variables like *Overall Impression* (R^2 = 0.34), *Concise*

Table 4.9 – Descriptive stats of social variables and nonverbal cues. Mean values for speaking cues are after z-score normalization.

	Mean Value			
Nonverbal cues	Session 1	Session 2	Delta	pValue
Overall Impression	5.2	5.5	0.3	0.03
Communicative	5.5	5.7	0.2	0.03
Sociable	5.2	5.4	0.2	0.02
Persuasive	5.2	5.5	0.3	0.01
Applicant Speaking Time	0.60	0.66	0.6	0.008
Applicant Turn duration	0.11	0.13	0.02	0.005

Table 4.10 – Regression results (N = 169) for verbal, nonverbal and combining both cues using RF ($p < 0.05$ for all).

Variables	Baseline	Nonverbal		Verbal		Nonverbal + Verbal	
	R^2	R^2	RMSE	R^2	RMSE	R^2	RMSE
motiv	0.0	0.26	0.46	0.13	0.54	0.27	0.46
compe	0.0	0.21	0.36	0.17	0.37	0.26	0.33
hardw	0.0	0.21	0.41	0.16	0.43	0.26	0.38
socia	0.0	0.27	0.45	0.03	0.59	0.22	0.48
enthu	0.0	0.33	0.70	0.04	0.98	0.32	0.69
posit	0.0	0.35	0.55	0.06	0.80	0.33	0.57
commu	0.0	0.26	0.52	0.04	0.67	0.26	0.52
conci	0.0	0.21	0.52	0.09	0.59	0.26	0.48
persu	0.0	0.33	0.67	0.08	0.92	0.32	0.67
ovImp	0.0	0.32	0.82	0.11	1.07	0.34	0.79

(R^2 = 0.26) and all the professional skills ($R^2 \in [0.26, 0.27]$) indicating that verbal components adds some information. The work in [112] investigated words used and nonverbal behavior displayed in a job interview setup using college students. They examined a different set of social variables and used correlation coefficient r as their evaluation measure. They too used LIWC for extracting lexical features but then applied LDA to learn common topics in the data corpus. By combining lexical and nonverbal features, the authors reported a prediction accuracy of r = 0.70 for *Overall Performance*, which indicates a R^2 = 0.49 compared to our R^2 = 0.34. This dataset is not publicly available (to the best of our knowledge) and thus, there is no direct way to assess the performance difference.

To understand the contributions of each of the feature sets, we determine the top 20 features used by RF model, presented in Table 4.11. We observe that while most of the features are nonverbal cues (*Speaking Energy, Turn Duration, Silence Events etc*), some verbal cues like *Question, Word Count, Proper pronouns* also contribute to the inference of *Overall Impression*, indicating that verbal cues add albeit marginally to improved inference.

4.5.4 Inferring Principal Components from Features

We define a second regression task with the aim of predicting the first three PCs which account for 94.3% of the variance in the annotation data. The best inference performance was achieved by using RF (Table 4.12). We observe that predicting the first PC using nonverbal cues achieves R^2 = 0.41 which is better than the performance for all the individual social variables. Essentially, this predicts positive or negative impression.

Similarly, using verbal components we can infer only up to 0.12. This suggests that use of PCs removes some of the noise in the annotations data leading to slightly improved inference. We also observe that the second and third PC are not recognizable, likely due to the little variance

Table 4.11 – List of top 20 features from RF regression model (N = 169) for *Overall Impression* (left:1 – 10; right:11 – 20)

Applicant Features	
Speaking Energy	Max Energy Derivative
Energy Derivative Lower Quartile	Min Energy Derivative
Avg Speaking Energy	Speaking Energy Lower Quartile
Avg Speaking Turn Duration	Speaking Energy Upper Quartile
Silence Ratio	Energy Derivative Upper Quartile
Number of Silence Events	Max Turn Duration
Max Silence Duration	Questions
Number of Speaking Turns	Word Count
Number of Head Nods	3rd Person Singular
Max Speaking Energy	Proper Pronouns

Table 4.12 – Inferring of principal components (PC) using random forest (RF) with verbal, nonverbal and their combining as predictors ($p < 0.05$).

Cues	PC1	PC2	PC3
Nonverbal	0.41	0.04	0.01
Verbal	0.12	0.02	0.02
Nonverbal + Verbal	0.34	0.02	0.02

(6.1% and 5.6% respectively) they account for.

4.6 Conclusion

In this chapter, we described a number of challenges associated with implementing a behavioral training procedure for hospitality students in order to improve the first impressions that others form about them. The framework consisted of two scenarios that were relevant for their future careers, namely job interviews and reception desk situations. We collected a new corpus of 169 simulated job interviews and reception desk interactions (338 interactions). This dataset was recorded with multiple modalities.

Specifically, this chapter analyzed the formation of first impressions in job interview situation and its connection to verbal and nonverbal cues displayed. Nonverbal cues were automatically extracted and their relationship with various perceived social variables analyzed. Our results are comparable with those reported in [114] and [112]. One of the insights from our analysis is that language has an effect on the predictive power of impressions scores. Using data from only French language interviews showed higher prediction accuracy ($R^2 = 0.32$) than interviews which were conducted in only English ($R^2 = 0.14$) for overall impression, while accuracy for other social variables (except “concise”) was comparable. Understanding these issues in more

depth requires future work.

Another insight was the role of gender in prediction accuracy. Interviews with male participant were predicted with higher accuracy ($R^2 \in [0.13, 0.46]$) than the ones featuring females ($R^2 \leq 0.12$). As the difference in N is relatively small (57% female to 43% male) and both data subsets have comparable size (or larger) compared to other existing datasets, we believe this is an interesting result. This result is supported by findings in psychology [90] and is in line with gender stereotypes for men, and with persuasiveness.

To understand the connection between linguistic content and impressions, the interactions were first manually transcribed. Then, verbal cues were extracted using LIWC, which links linguistic and paralinguistic categories to psychological constructs. A correlation analysis between use of words and impression scores provided interesting insights into the weak effect of linguistic content on impressions, a result in line with some existing literature. An inference of impressions scores defined as a regression task showed that verbal features had low performance as compared to nonverbal cues, indicating the importance of the latter in a structured job interview context.

We then assessed the underlying structure of the annotations using principal component analysis. The first PC accounted for more that 82% of the variance and was found to distinguish the overall impression. Using this PC as labels in a regression task showed performance of $R^2 = 0.41$. This implies (1) the use of PCs removes some of the noise in the annotations data leading to slightly improved inference (2) there exists a lower dimensional representation of skills impressions.

5 First Impressions in Reception Desk

The interaction between service employees (e.g. reception desk assistants) and customers, commonly referred to as service encounters in the hospitality industry, is a critical part of customer experience at an establishment [157]. It is during these encounters that customers perceive and evaluate the employee's attitudes and professional, social, and communication skills. Based on these interactions, customers form impressions of both the employee and the organization [103]. The importance of interpersonal communication during service encounters in determining customer satisfaction and perceived quality of service (QoS) has been highlighted in prior work [55, 157]. Literature in psychology, marketing, and hospitality has demonstrated that as a major component of interpersonal communication [87], nonverbal behavior (NVB) contributes towards shaping the outcome of customer-employee interactions [54, 157]. Customers often use interactions with front line service employees to assess QoS [67], it is imperative for hospitality organizations to improve the quality of these encounters. In this work, bringing audio-visual processing and machine learning as additional analytical tools, we investigate the connections between automatically extracted nonverbal behavior and performance impressions in service encounters, in addition to other important factors including employees' personality traits and attractiveness. Specifically, we study dyadic interactions at a hotel reception desk setting between employees and customers.

Job performance is a central construct in organizational psychology and has a variety of definitions in the literature [163]. Specific aspects of the performance construct may change from job to job, but some dimensions can be generalized across jobs (e.g. interpersonal communication). In this work, we use the definition proposed in [163], which denotes job performance as "action, behavior and outcomes that employees engage in and contribute to organizational goals". We define *performance impressions* as the behavioral aspect of performance as perceived by others who can observe an interaction (e.g. in a dyadic service encounter) and assess the performance of the employee based on the interaction itself.

In this chapter, we study the connections between performance and related impressions and automatically extracted nonverbal behavior, as well as other relevant variables discussed in hospitality, marketing, and psychology literature, namely perceived personality traits and

attractiveness. Our work builds upon and extends work on automatic analysis of social interaction in the workplace [58, 130], which has shown the potential of inferring negotiation outcomes [36], job interview ratings [114, 28], and other constructs (e.g. engagement, friendliness, or excitement) [75, 141, 112] up to a certain level of performance. We study a dataset consisting of 169 dyadic interactions between a hotel desk receptionist and a client, where receptionists are played by students from an international hospitality management school who practice real-life situations. We address the following research questions:

RQ1: What nonverbal cues displayed by desk service employees are connected to performance and skill impressions? Can they be used to automatically infer these constructs?
RQ2: How are perceived Big-5 personality traits of desk service employees linked to performance impressions?
RQ3: Are there any connections between the perceived attractiveness of such employees and their performance impressions?

The contributions of this work are the following. First, we analyze the relationship between automatically extracted audio-visual nonverbal cues and performance and skill impressions via correlation analysis and a regression task. Interestingly, we show that the customer's nonverbal cues explain up to 27% of the variance of the participant's perceived performance scores. Second, we show that Big-5 trait impressions achieve performance of $R^2 = 0.35$ for job performance impressions. Third, we show that judgments of attractiveness were not good predictors of impression and skill ratings. Finally, the integration of NVB cues and Big-5 impression results in an overall benefit in inference of performance impressions. This research might have wider implications for employees and managers in hospitality, by providing an understanding of not only the employee's performance via nonverbal behavior, but also of the implications for customer perceptions of service encounters. The automatic approach could also facilitate personalized training for employees to improve their nonverbal behavior in service encounters.

The material of the chapter was originally published in [110].

5.1 Dataset

The reception desk is considered the entry point of a hotel and, as a very frequent type of interaction in hospitality, is often determinant of the evaluations of service quality of such organizations [138, 67]. Despite its importance, there is no publicly available dataset to study this interaction from the perspective of performance impressions. We used a data corpus consisting of 169 interactions in a reception desk setting, described in Section 3.2. Here we again briefly outline the desk situation and annotation of attractiveness, impressions of performance, skills and personality traits.

5.1.1 Data collection & Annotations

The reception desk role-play involves a receptionist (a hospitality student participant) and a client (a research assistant selected from a seven-person group of master's students in business and psychology). Study participants were students at an international hospitality management school. A total of 100 students voluntarily took part in the study (mean age 20.6 years old; 57 females and 43 males). 69 participants contributed two reception desk interactions. Video data was collected with two Kinect v2 devices (one for client, one for receptionist), each recording 30 fps RGB+depth video (1920×1080 and 512×424 for RGB and depth, respectively.) Audio data was collected using a microphone array device, that captures audio at 48kHz and segments speaker turns from localized sources. The audio and video streams are synchronized in a subsequent step.

The reception desk videos were annotated by two separate group of raters. Group-A coded impressions of performance, skills and personality traits (Big-5), while (Group-B) codded attractiveness attributes. Details of the annotation task can be found in Section 3.3. The descriptive statistics and the agreement between raters measured using the Intraclass Correlation Coefficient (ICC) (a commonly used metric in psychology [147]) is summarized in Table 6.1. We observe that the agreement between raters for performance and skill impression variables was moderate to high, with $ICC(2,k) \in [0.58, 0.77]$. While, for personality traits impression $[0.41 < ICC(2,k) < 0.68]$ and attractiveness attributes $[0.36 < ICC(2,k) < 0.62]$ the agreement between raters was moderate. For Big-5, this could be due to the interaction setting, which elicits *Agreeableness* and *Extraversion* traits to be more visible.

5.2 Verbal and Nonverbal Features

To understand the influence of nonverbal behavior on the formation of performance and skill impressions, various cues were automatically extracted from the audio and visual modalities from both the receptionist and client. The complete description of the verbal and nonverbal features is presented in Section 3.5, we list them here briefly for sake of completeness (Table 5.2). The choice of nonverbal and verbal cues were guided by their relevance in existing literature in social psychology [40, 77, 6], hospitality [55, 157] and social computing [92, 114, 108]. The nonverbal cues were extracted from the moment the client gets the bill and changes to an unfriendly attitude until the end of the interaction. The reception desk videos were first manually transcribed (Section 3.4). Then, these transcripts were processed to extract lexical features using the Linguistic Inquiry and Word Count (LIWC) software [127], widely used in social psychology [32] and social computing [20, 143] to extract verbal content.

5.3 Correlation Analysis

This section presents the Pearson correlation analysis performed to understand performance and skill impressions in this setup and their relationship with nonverbal cues, personality

Table 5.1 – Reception desk dataset: $ICC(2,k)$ & descriptive statistics for impressions of skills and performance, attractiveness & personality traits.

Variable	**ICC**	**mean**	**std**	**median**	**skew**
Professional Skills					
Competent (compe)	0.69	4.24	1.36	4.33	-0.30
Motivated (motiv)	0.63	4.80	1.12	5.00	-0.46
Satisfying (satis)	0.73	4.16	1.41	4.33	-0.15
Social Skills					
Intelligent (intel)	0.58	4.52	1.04	4.67	-0.18
Positive (posit)	0.60	4.34	1.09	4.33	-0.07
Sociable (socia)	0.64	4.46	1.14	4.33	-0.26
Communication Skills					
Clear (clear)	0.66	4.56	1.25	4.67	-0.53
Persuasive (persu)	0.72	4.01	1.38	4.00	-0.07
Overall					
Overall Impression (ovImp)	0.75	4.27	1.46	4.33	-0.13
Performance (peImp)	0.77	4.11	1.37	4.33	-0.06
Big-5 Personality					
Agreeableness (agree)	0.62	3.5	0.59	3.5	0.14
Conscientiousness (consc)	0.41	3.9	0.30	4.0	-0.44
Extraversion (extra)	0.68	4.23	0.41	4.33	-0.14
Neuroticism (neuro)	0.47	4.11	0.38	4.0	0.35
Openness (open)	0.40	4.19	0.29	4.17	0.71
Attractiveness					
Attractiveness (attrac)	0.62	3.72	1.44	3.73	0.27
Dislikeable (disli)	0.36	3.96	1.26	4.01	-0.18
Friendly (frien)	0.59	3.77	1.43	3.41	0.14
Likeable (likea)	0.55	3.48	1.35	3.41	0.21

Table 5.2 – Nonverbal features extracted from the reception desk data.

Features			
Speech Activity	Prosody	Visual	Multimodal
Speaking time	Pitch	Overall visual motion	Speaking while nodding
Speaking turns	Speaking Rate	Head nods	
Pauses	Spectral Energy	Visual Back-channelling	
Short Utterance	Speaking Energy		
Silence	Voicing Rate		
	Rate of change of energy		

Table 5.3 – Correlation between Big-5 personality trait impressions and *Performance Impressions* ($N = 169$; $^{*}p < 0.001$, $^{\dagger}p < 0.01$, $^{**}p < 0.05$). Entries without p-value symbol are not statistically significant.

Impressions	2	3	4	5	6
1.peImp	$.23^{\dagger}$	$-.11$	$.42^{*}$	$-.39^{*}$	$.15^{**}$
2.agree		$-.05$	.08	$-.45^{*}$	$-.01$
3.cons			$-.05$	.05	$-.03$
4.extra				$-.02$	.08
5.neuro					$-.08$
6.open					

trait impressions, and attractiveness. For the analysis, the average of each impression variable provided by the three raters is used.

5.3.1 Performance Impressions & Personality Trait Impressions

The correlation between Big-5 personality impressions and *performance impression* was computed (Table 5.3). *Extraversion* was observed to be positively correlated to *performance impression* with $r = 0.42$ ($p < 0.001$), while *Neuroticism* was found to be negatively correlated with $r = -0.39$ ($p < 0.001$). *Agreeableness* and *openness* were observed to have lower correlation to *performance impression* with $r = 0.23$ ($p < 0.01$) and $r = 0.15$ ($p < 0.05$) respectively. Interestingly, we do not observe any correlation between *Conscientiousness* and *performance impression* as suggested in [12]. This could be explained as in this situation, *Conscientiousness* is a hard trait to score and has lower agreement among raters ($ICC(2,k) = 0.41$). The results of other personality trait impressions are in line with literature in psychology, especially *Extraversion*, which is reported to be a valid predictor of performance for jobs requiring social interactions [12].

5.3.2 Performance Impressions & Nonverbal Cues

Receptionist

In the next step, correlations between extracted nonverbal cues and annotated variables were investigated. Correlation of selected nonverbal cues are presented in Table 6.7. A number of receptionist's features were found to be significantly correlated to impressions of performance and skills. Specifically, receptionists who spoke for longer duration, faster, took longer turns, and had fewer silence events obtained higher scores for performance and skill impressions. Similarly, receptionists who spoke animatedly with higher visual motion, nodded more, displayed greater visual BC, were more favorably viewed than those who spoke with less visual activity. Literature in psychology and hospitality has reported that the use of faster speech, with fewer silent events, enhances customer's perception of competence, while more head nods and visual BC enhances the perception of empathy, courtesy, and trust [157]. Our re-

Table 5.4 – Pearson correlation between nonverbal cues and performance and skill impressions $N = 163$;* $p < 0.001$,† $p < 0.01$, ** $p < 0.05$.

NVB Cues	Skills			peImp
	Professional	Social	Communication	
Receptionist				
Speaking Activity				
Speaking Ratio	[.35, .43]†	[.37, .44]†	[.30, .39]†	.43†
Mean Turn Duration	[.32, .37]†	[.33, .38]†	[.30, .34]†	.40†
Max Turn Duration	[.39, .41]†	[.36, .39]†	[.34, .41]†	.42†
Num Silence Events	[−.23, −.22]†	[−.29, −.18]†	[−.21, −.18]†	−.22†
Voicing Rate				
Mean	[.31, .34]†	[.32, .34]†	[.32, .32]†	.28†
Voicing Rate Q25	[.29, .32]†	[.28, .35]†	[.28, .29]†	.28†
Voicing Rate Q75	[.27, .30]†	[.27, .35]†	[.28, .31]†	.24†
Visual Motion				
Mean WMEI	[.19, .33]†	[.18, .36]†	[.20, .22]†	.26†
Max WMEI	[.30, .33]†	[.26, .37]†	[.31, .31]†	.30†
Count Head Nods	[.35, .42]†	[.39, .41]†	[.34, .41]†	.37†
Visual BC				
Count	[.20, .29]†	[.26, .27]†	[.22, .25]†	.23†
Mean Duration	[.22, .22]†	[.20, .25]†	[.20, .20]**	.18**
Max Duration	[.25, .30]†	[.26, .29]†	[.23, .25]†	.25†
Min Duration	[−.26, −.19]†	[−.26, −.18]†	[−.24, −.19]**	−.22†
Multimodal Cues				
Count	[.44, .49]†	[.45, .49]†	[.41, .49]†	.45†
Mean Duration	[.23, .30]†	[.22, .27]†	[.26, .30]†	.24†
Max Duration	[.40, .43]†	[.40, .43]†	[.39, .44]†	.39†
Min Duration	[−.27, −.23]†	[−.33, −.25]†	[−.24, −.23]†	−.24†
Client				
Voicing Rate				
Mean Voicing Rate	[.24, .31]†	[.30, .35]†	[.23, .24]†	.25†
Voicing Rate Q25	[.17, .23]†	[.23, .26]†	[.19, .21]†	.19**
Voicing Rate Q75	[.24, .27]†	[.27, .29]†	[.21, .22]†	.21†
Visual Motion				
Max WMEI	[.30, .33]†	[.26, .37]†	[.31, .31]†	.33†
Count Head Nod	[.18, .30]†	[.24, .29]†	[.27, .31]†	.24†
Visual BC				
Count	[.25, .30]†	[.28, .30]†	[.26, .31]†	.30†
Max Duration	[.20, .21]†	[.22, .24]†	[.17, .21]†	.17**
Min Duration	[−.24, −.22]†	[−.20, −.16]†	[−.22, −.17]†	−.21†

Table 5.5 – Range of Pearsons correlation between eye gaze, facial expressions and social variables in the reception desk ($N = 153$) dataset. Due to nonfrontal face in this setting, the N is different from Tabl 6.7. $^{***}p < 0.001$;$^{**}p < 0.01$;$^{*}p < 0.05$.

	Professional	Social	Communication	Performance
Receptionist				
Num of GWS	[.27, .34]***	[.26, .32]***	[.24, .28]***	.30***
Max Duration (GWS)	[.14, .21*]	[.21, .22]**	[.22, .24]**	.20*
Min Duration (GWS)	[.17, .18*]	[.17, .19]*	[.15, .20*]	.21*
VDR	[.31, .32]***	[.28, .32]***	[.24, .31]***	.36***
Facial Expression				
Std Anger	[.21, .27]**	[.16, .20*]	[.20, .21]*	.22**
Var Anger	.21*	[.16, .19*]	[.19, .19]*	.22**
Max Anger	[.22, .26]**	[.18, .20]*	[.20, .22]**	.24**

sults are comparable with previous literature for other conversational settings like interviews [114, 40], restaurant service [81], and sales [6].

We observed moderate correlations between eye gaze, facial expressions and perceived soft skills (Table 5.5). In particular, we observe that participants who held client's gaze for longer duration while speaking, displayed greater VDR and nonverbal immediacy (by mirroring anger the clients displayed), were perceived to be better performing then participants who did not display these cues. These connections are supported in literature. It has been shown that even in situations when the client is dissatisfied, greater eye contact leads to enhanced perception of credibility [157]. Soderlund et al. [152] reported that dissatisfied client's assessment of employee's emotional state affects their own emotional state, which, in turn, impacts customer's level of satisfaction.

Client

An interesting insight is the correlation between some of the client's nonverbal cues and the impression score of the receptionist. We observed that clients tend to speak faster with greater visual motion in presence of receptionists who were rated higher. Also, clients tend to nod more and provide greater visual BC and for longer duration while interacting with receptionists who scored higher than with receptionists with lower scores. Similar results are reported in other dyadic settings like job interviews [114, 108].

5.3.3 Performance Impressions & Attractiveness

The correlation analysis of attractiveness attributes and *performance impression* yielded unexpected results (Table 5.6). It was observed that *attractive* was not significantly correlated to *performance impression*, while *friendly* ($r = -0.27$) and *likeable* ($r = -0.26$) had low negatively correlation ($p < 0.05$). *Dislikeable* was observed to have low positive correlation ($r = 0.17$; p

Table 5.6 – Correlation between *Attractiveness* attributes and *Performance Impression* (N = 163;* $p < 0.001$,† $p < 0.01$, ** $p < 0.05$). Entries without p-value symbol are not statistically significant.

	All Receptionists				Female Receptionists				Male Receptionists			
Impressions	2	3	4	5	2	3	4	5	2	3	4	5
1.peImp	−.12	.18**	−.27**	−.26**	−.18	.40*	−.47*	−.48*	−.02	−.05	.08	−.06
2.attra		−.44*	.60*	.54*		−.56*	.68*	.69*		−.26**	.47*	.33†
3.disli			−.74*	−.78*			−.86*	−.85*			−.59*	−.72*
4.frien				.83**				.89*				.75*
5.likea												

< 0.05). Given the literature on gender and attractiveness and job performance [151, 80], we divided the sample of receptionists based on gender. It was observed that for males (N = 79) there was no correlation between any attractiveness attributes and *performance impression* ($r \in [0.01, -0.05]$). However, for the female receptionists (N = 90), *friendly* ($r = -0.47$) and *likeable* ($r = -0.48$) was negatively correlated to *performance impression* ($p < 0.001$), while *dislikeable* was positively correlated ($r = 0.40$; $p < 0.001$). This result does not conform several of the results reported in the literature of attractiveness and performance, where a positive connection was often found [66, 151]. For further discussion, refer to Section 5.4.3.

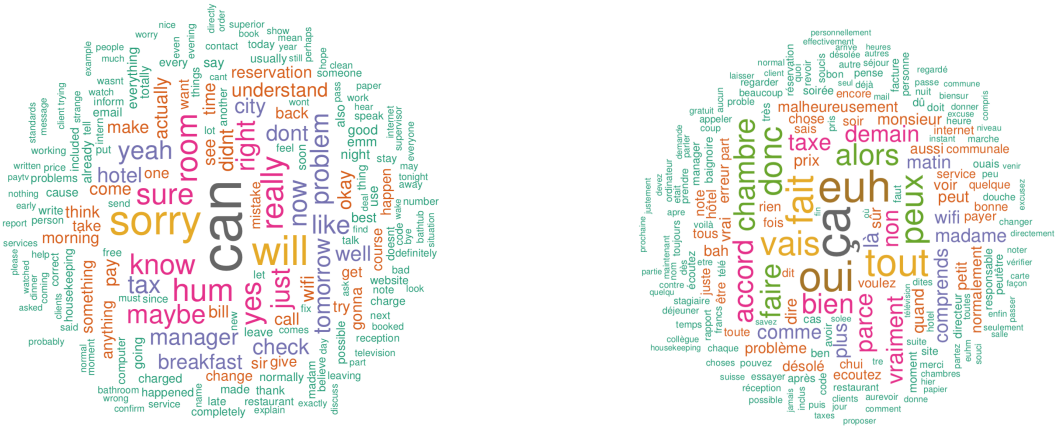
5.3.4 Performance Impressions & Verbal Cues

As a first step towards understanding the connections between choice of words during the desk interactions, and performance and skill impressions, we generate a wordcloud to visualize the most frequently occurring words. This wordcloud was generated in Python using pandas and matplotlib from the desk transcripts. Note that the stopwords were removed in both languages before generating the wordcloud. From Figure 5.1a, 5.1b, we observe that the most commonly used words during the English interactions are *can, sorry, will, really, room, sure, yes, know, and right*, while for the French interactions are *ca, oui, tout, fait, vais, peux, donc, alors, and chambre*. Most of the words are specific to the context of the desk interaction.

In the next step, we conduct a Pearson's correlation analysis of LIWC features with performance and skill impressions (Table 5.7). For this analysis, the mean rating of all variables by the five raters and the features extracted from LIWC were utilized. Overall, we observe low correlation between verbal content and *Performance Impression*. We observe that participants who spoke more (greater word count), used speech containing 3rd person plural (*they, their*), future tense (*will, going*), and negative emotions (*hurt, angry*) prepositions were rated more positively. Furthermore, use of 3rd person singular (*she, him*), negation (*no, not*), and certainty (*always, certainly*) are negatively correlated to performance impressions.

5.4 Regression Analysis

A framework for inference of impressions of performance and skills from nonverbal cues, personality impressions, and attractiveness impressions was proposed and evaluated. The



(a) (b)

Figure 5.1 – Wordcloud showing the frequency of words used during the 169 desk interactions. In this figure, the relative size of each word indicates its frequency. For example, in the English interactions, the most common word is *can*, while for French interactions it is *ca*. All words of the same color have the same size/frequency.

Table 5.7 – Pearson correlation between nonverbal cues and performance and skill impressions $N = 163$; $^{*}\,p < 0.05$, $^{**}\,p < 0.01$, $^{***}\,p < 0.001$

Impressions	WC	3rd person singular	3rd person plural	future tense	preposition	negations	humans	negative emotions	certain
Performance	0.40***	-0.22*	0.25**	0.24**	0.27**	-0.21*	-0.23**	0.22**	-0.23**
Competent	0.40***	-0.17*	0.22**	0.19*	0.24**	-0.22*	-0.27**	0.16	-0.21*
Motivated	0.43***	-0.15	0.13	0.18*	0.30***	-0.07	-0.17*	0.11	-0.23**
Satisfying	0.32***	-0.18*	0.26**	0.25**	0.21*	-0.22**	-0.26**	0.20*	-0.20*
Intelligent	0.38***	-0.17*	0.18*	0.21*	0.24**	-0.25**	-0.28***	0.19*	-0.16
Positive	0.40***	-0.1	0.30***	0.16	0.19*	-0.15	-0.28***	0.13	-0.18*
Sociable	0.44***	-0.11	0.26**	0.14	0.20*	-0.07	-0.20*	0.09	-0.17*
Clear	0.36***	-0.20*	0.17*	0.16	0.20*	-0.19*	-0.29***	0.17*	-0.22**
Persuasive	0.36***	-0.18*	0.26**	0.21*	0.20*	-0.16	-0.24**	0.22*	-0.21*

data was first preprocessed by a person-independent Z-score normalization to transform data into unity variance and zero mean. Then, both a full feature representation and standard dimensionality reduction techniques (Principle Component Analysis and significantly correlated features) were evaluated.

Regarding the machine learning approach, two regression techniques (Ridge Regression (Ridge) and Random Forest (RF)) implemented in the Caret R package were evaluated [88]. Leave-one-person-out cross-validation and 10-fold inner cross-validation were used. Hyper parameters (i.e., number of trees, shrinkage parameters) were automatically tuned by using an inner 10-fold cross-validation on the training set. Performance of these regression techniques were evaluated by employing two standard measures: coefficient of determination (R^2) and root-mean-square error ($RMSE$). Here, results of only the best performing model are presented and discussed. For this task, as the baseline we use $R^2 = 0.0$ by predicting the population mean.

Table 5.8 – Best inference performance results using NVB cues, personality traits (Big-5) impressions, Attractiveness impressions and various combination of impressions and NVB. All results were significant with $p < 0.05$ ($N = 169$). Best performing model is indicated by * (RF); ** (Ridge)

Impressions and Skills		Baseline		Nonverbal*		Big-5*		Attract**		NVB + Big-5*		NVB + Attract**	
		R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE
Performance	peImp	0.0	0.0	.30	1.31	.35	1.22	.18	1.30	.37	1.18	.21	1.31
Professional	compe	0.0	0.0	.29	1.34	.30	1.33	.15	1.32	.36	1.18	.17	1.33
	motiv	0.0	0.0	.30	.88	.29	.89	.14	1.02	.34	.83	.12	1.03
	satis	0.0	0.0	.29	1.42	.32	1.36	.16	1.31	.36	1.28	.13	1.30
Social	intel	0.0	0.0	.26	.81	.27	.82	.13	1.08	.28	.80	.11	1.07
	posit	0.0	0.0	.32	.79	.33	.78	.12	1.06	.41	.69	.17	1.06
	socia	0.0	0.0	.33	.85	.43	.73	.13	1.07	.44	.71	.12	1.07
Communication	clear	0.0	0.0	.22	1.21	.25	1.16	.15	1.26	.28	1.13	.14	1.26
	persu	0.0	0.0	.32	1.30	.29	1.36	.15	1.28	.36	1.22	.14	1.28

5.4.1 Inferring Performance Impressions from Nonverbal Cues

Regression results indicate that all variables could be predicted to a certain degree from automatically extracted nonverbal cues (Table 5.8). It is observed that 30% of variance in *performance impression* (peImp) can be explained by nonverbal cues. Other variables have similar predictability using aggregated nonverbal cues of *Participant* and *Client*. Specifically, *sociable* (socia) has the highest performance ($R^2 = 0.33$), followed by *positive* (posit), *persuasive* (persu) (both $R^2 = 0.32$), and *motivated* (motiv) ($R^2 = 0.30$). These results provide an answer to RQ1: nonverbal behavior is predictive of performance and skill impressions in this hospitality encounter scenario. These results also corroborate findings in other conversational settings like job interviews [114, 108] and job negotiations [36]. In [36] the authors were able to explain up to 30% of the variance in job negotiation performance using audio features, while the authors in [114, 108] reported $R^2 = 0.34$ and $R^2 = 0.32$ for hirability and overall impressions respectively. Our results are in the same range. In hospitality literature, it has been shown that nonverbal behavior is correlated to customer satisfaction ($r \in [0.33, 0.42]$) [81]. We compare these results to this work by converting r to R^2 (our evaluation measure, coefficient of determination R^2, is approximated by computing the square of correlation coefficient r). They reported a prediction accuracy of $r = 0.42$ for overall performance, which indicates a $R^2 = 0.18$. Some variables like *clear* (clear) ($R^2 = 0.22$) are harder to predict using extracted nonverbal cues. Our results can be seen as baseline for this type of task in the reception desk setting, and the reported R^2 is comparable with results obtained in other tasks in the literature.

We observe that eye gaze has low inference performance with the best $R^2 = 0.10$ for *Positive* (Table 5.9). Facial expressions too had low inference capability with the best $R^2 = 0.08$ for *Positive*. Combing these two cues showed improved inference performance. In particular, the combined set of features had a best performance with $R^2 = 0.15$ for *Positive*, *Performance*, while we also observed improved performance for *Motivated* ($R^2 = 0.13$) and *Persuasive* ($R^2 = 0.12$).

Table 5.9 – Regression analysis using eye gaze, facial expressions and combination of features for desk ($N = 153$) setting.

	Clear	Persuasive	Positive	Social	Competent	Motivated	Performance
Gaze	0.00	0.02	0.10	0.01	0.06	0.09	0.08
Facial Expressions	0.00	0.04	0.08	0.05	0.00	0.03	0.01
Gaze + Expressions	0.05	0.12	0.15	0.10	0.07	0.13	0.15
All Visual	0.21	0.31	0.32	0.30	0.31	0.25	0.30
All Features	0.24	0.33	0.34	0.32	0.31	0.29	0.32
Baseline	0.22	0.32	0.32	0.33	0.29	0.30	0.30

Inference performance improve further when gaze and facial expressions were fused with other visual features. The best performance was obtained for *Positive* ($R^2 = 0.32$) followed by *Persuasive* and *Competent* ($R^2 = 0.31$). This fused set of features obtained $R^2 = 0.30$ for *Performance* and *Social*. As a last step, we fused all the visual features with auditory behavioral cues. We observe a moderate improvement in inference performance (as compared to the baseline) for all the variables except *Social* and *Motivated*. Specifically, we observe improved inference for *Positive* ($R^2 = 0.34$), *Persuasive* ($R^2 = 0.33$), *Social* and *Performance* (both $R^2 = 0.32$).

These results are in accordance with those reported in literature. Leigh et al. [91] evaluated impact of eye gaze and other nonverbal behavior on perceived performance of salespersons. They reported significant effect of eye gaze on perceived believability,tactfulness and empathy. DeGroot et al. [40] investigating perceived performance of 110 managers in a news-publishing company, reported a correlation of $r = 0.14$ between visual cues displayed and performance.

As a next step, the contribution of the nonverbal cues from each protagonist was investigated. Receptionist's cues contribute to the predictive performance of all variables with $R^2 = 0.28$ for *performance impression*. An interesting result is that client's nonverbal cues explains variance in *performance impression* ($R^2 = 0.27$) almost as much as receptionist's own nonverbal cues. Similar results are observed for other skill impressions and are analogous to results reported in [114], where nonverbal cues of the interviewer contributed ($R^2 = 0.22$) to explaining the variance in applicant's hiring scores. The effect of gender on predictive power of nonverbal cues was investigated. The dataset was divided based on gender and regression experiments were rerun. No major difference in predictive performance of nonverbal cues was observed, with $R^2 = 0.28$ for female and $R^2 = 0.27$ for male participants for *performance impression*.

5.4.2 Inferring Performance Impressions from Big-5 Impressions

To investigate the role of personality impressions predicting *performance impression*, a regression task was defined with the personality impressions as predictors. It is observed that the RF model explains up to 35% of the variance in the data. Similarly, these trait impressions performed moderately for other impressions, with highest R^2 achieved for *sociable* (0.43). Overall, all performance and skill impressions have moderate predictability [$R^2 \in (0.25, 0.43)$] using Big-5 impressions as predictors and finds support in the literature [12], answering RQ2 and validating the predictive power of Big-5 trait impressions in service encounters. As a next

step, we combined the personality trait impressions and automatically extracted nonverbal cues to infer *performance impression*, and achieved $R^2 = 0.37$, which is marginally higher than each single source of information. In the case of *sociable* , up to 44% of variance in impression scores could be explained (highest among all skills). Overall, the highest performance for the inference task was achieved by combining nonverbal cues and Big-5 impressions implying that these impressions added extra information to the NVB cues.

5.4.3 Inferring Performance Impressions from Attractiveness

To further analyze the link between attractiveness variables and *performance impression*, a regression task was defined with the aim of evaluating the predictive power of attractiveness attributes as predictors. The ridge regression model performed best for this case and the results are presented in Table 5.8. From the table, it is observed that $R^2 < 0.20$, indicating that attractiveness variables had low predictive power. A similar observation was made while analyzing attractiveness attributes derived from still images in interview context by [114]. This result perhaps could be explained by the fact that raters annotated attractiveness variables by looking at still images rather than video clips. The methodology of using still images for attractiveness annotation, though has solid backing in psychology literature [8, 66], did not produce positive results in our case. In other works, authors reported a positive effect of physical attractiveness on performance impressions of sales representatives [2], teachers [8], and service workers [151]. This issue has to be investigated further.

In the computing literature using video instead of still images, Biel et al. in [19] annotated two facets of physical attractiveness, and three facets of non-physical attractiveness on 442 1-min YouTube videos and reported that more attractive people were often judged as having more positive traits. Specifically, the authors reported correlation between *Agreeableness* and *Friendliness* to be $r = 0.57$ ($p < 0.001$). In this work, we found that the correlation between traits like *Agreeableness* (labeled on videos) and attractiveness variables like *Likeable* (labeled on images) to be very low ($r = 0.0.07$).

A hypothesis for this weak connection might be due to the difference in the amount of perceptual cues available for *performance impression* (video) and perceived attractiveness (still images). This hypothesis might find some support in [24], which reported that while a still image was a valid modality to infer various personality traits, a greater validity was achieved using audio-visual clips. The authors also state that “there are relations between physical attributes and personality traits, and subjects are quite aware of these relationships”. For RQ3, we conclude that attractiveness variables inferred from still images have little connection to *performance impression* in our specific setting. In future work, we plan to investigate the use of video data for annotations of perceived attractiveness and its connections to performance and skill impressions.

Table 5.10 – Inference results using verbal cues, fusion of verbal cues with NVB cues and personality traits (Big-5) impressions. All results were significant with $p < 0.05$ ($N = 169$). Best performance is obtained using RF model.

Impression and Skills		Verbal		Verbal + NVB		Verbal + NVB + Personality	
		R2	RMSE	R2	RMSE	R2	RMSE
Performance	peImp	0.23	1.66	0.32	1.5	0.41	1.29
Professional	compe	0.18	1.53	0.30	1.31	0.35	1.21
	motiv	0.10	1.2	0.24	0.97	0.28	0.92
	satis	0.16	1.7	0.28	1.45	0.34	1.31
Social	intel	0.19	0.90	0.28	0.79	0.31	0.76
	posit	0.19	0.97	0.33	0.81	0.41	0.7
	socia	0.18	1.01	0.36	0.79	0.41	0.72
Communication	clear	0.16	1.3	0.28	1.11	0.29	1.11
	persu	0.15	1.65	0.29	1.36	0.34	1.27

5.4.4 Inferring Performance Impressions from Verbal Cues

In the last step, we study the feasibility of inferring all the variables of interest using verbal cues (Table 5.10). We obtain a best inference results for *performance impression* (R^2=0.23), followed by *Intelligent (intel)* and *Positive (posit)* (R^2=0.19), while verbal cues have low inference for *Motivated (motiv)* (R^2=0.10). Overall, verbal cues have low inference performance. These results are better than those reported in literature for those reported in other settings like inference of hirability in job interviews [106, 28], leadership [143], mood [142] and personality [20].

Combining verbal cues with NVB cues improves the inference performance of our RF model. Specifically, we see improved inference for all variables except *Motivated (motiv)* and *Persuasive (persu)*. The highest improvements is for *Clear (clear)* (R^2 = 0.28), followed by *Social (socia)* (R^2=0.36). We see improved inference for *Performance (peImp)* (R^2 = 0.32), which indicates that in reception desk setting, words used play a role in perception of job performance. The best inference performance for *performance impression* is obtained by fusing verbal cues, nonverbal cues and personality traits (Big-5) impressions (R^2 = 0.41). Overall, our experiments indicate that verbal cues capture some variance in perceived variable scores and contributes to improved inference for impressions of skill and performance.

5.5 Conclusion

This chapter described our investigation of the interplay between nonverbal behavior cues, Big-5 personality trait, and attractiveness impressions in hospitality service encounters, a novel setting in multimodal interaction research. We extracted a number of relevant verbal and nonverbal cues automatically and studied their relationship with perceived performance

and skill impressions.

We found that receptionists who spoke faster, for greater duration, took longer turns, and had fewer silence events had high scores for performance and skill impressions. These results are supported by literature in marketing [157].The inference task with NVB as predictors explains up to 30% of variance, and is comparable with results obtained for similar dyadic conversation settings in the literature.

We found that Big Five personality trait impressions are predictors of performance and skill impressions. Specifically, receptionists who conveyed higher *Extraversion* were rated higher in terms of performance, while receptionists who were high in *Neuroticism* were rated lower with respect to performance. This is in line with work on Big-5 and job performance in psychology [12], and extends the findings to the hospitality reception desk scenario. An inference task with Big-5 impressions as predictors achieved a performance of $R^2 = 0.35$, while integrating NVB cues and Big-5 trait impressions results in slightly improved performance ($R^2 = 0.37$).

Our work found a negative correlation between attractiveness attributes like *likeable* and *friendly* and *performance impression* scores for women participants, while there was no correlation for men. Extending this further into a regression task using attractiveness attributes as predictors, we observed low predictive power ($R^2 < 0.18$) for all performance and skill impressions.

We also investigated the connections between linguistic content and first impressions in this novel workplace setting. Towards this, the interactions were first manually transcribed. Then, we extracted verbal cues using LIWC software. A correlation analysis between use of words and impression scores revealed low to moderate correlations between linguistic content on impressions. An regression framework to infer impression scores showed that verbal features had lower performance compared to nonverbal cues, indicating the importance of the latter during a dyadic interaction with an unhappy client. Fusion of verbal and nonverbal cues improved inference performance slightly.

Finally, given the importance of service encounters on customer evaluation of quality of service, it seems essential for managers and employees to better understand how behavior might influence customer perception. Hence, this work could have implications for training and development of service employees. In the future, we plan to explore other behavioral cues including smiling, gaze, verbal content, and emotion recognition as features. We also plan to incorporate the findings of this work into a feedback system which provides automatic real-time feedback based on employee behavior.

6 Cross-situation analysis

One of the objectives of this thesis is to examine human behavior and its impact on impressions formed in two different situations. Previous chapters of this thesis investigated verbal and nonverbal behavior of participants in job interviews (Chapter 4) and reception desk (Chapter 5) individually. In this chapter, we investigate the connections between nonverbal behavior, verbal content, and first impressions in two different workplace situations: interviewing for a job and performing at the job.

People are known to behave differently in diverse situations, as person and situation are intricately entwined. Also known as "person-situation debate" or "person-situation-behavior triad", this has been a research topic in social sciences for decades [83]. Yet, until the advent of ubiquitous computing technologies, it had been difficult to objectively quantify behavior in multiple person-situation cases, due to lack of access (both direct and unobtrusive) to interactions across situations [51, 100]. Motowildo et al. studied aural and visual sources of nonverbal behavior and their correlations to performance on the job in a dataset consisting of 40 managers [104]. Similarly, DeGroot et al. evaluated the relationship between interviewees' nonverbal (visual and aural) and (a) impressions formed by the interviewers (b) interviewees job performance [40]. In both of these investigations, supervisors' ratings were considered as the measure of job performance.

In this chapter, we study the connections between first impressions and automatically extracted verbal and nonverbal behavioral cues from two different situations. Specifically, we investigate connections between perceived hirability and soft skills from job interviews, behavioral cues (verbal and nonverbal) displayed during both job interviews and the reception desk, and perceived job performance and soft skills from reception desk interactions. We define *perceived performance* as the behavioral aspect of performance as perceived by others observing an interaction (like a hotel front desk, or a sale) and assessing the performance of the employee based on the interaction itself. While job performance has varied definitions in literature, our definition is derived from that proposed by Viswesvaran et al. [163], who defined job performance as "action, behavior and outcomes that employees engage in and contribute to organizational goals". We note that while specific expressions of job performance depend

on the jobs and positions, some aspects can be generalized across jobs like interpersonal communication.

Towards this objective, we use a data corpus consisting of 338 videos of job interviews and reception desk interactions played by a sample of students from an international hospitality school (Chapter 3). We address the following research questions:

> **RQ1:** What are the connections between perceptions of candidates in job interviews and perceptions of the same person on the job?
> **RQ2:** What is the link between automatically extracted nonverbal behavior of candidates during job interviews and the perception of performance on the job?
> **RQ3:** What are the connections between candidates' choice of words in the two interactions and the perception of performance on the job?

To answer these questions, we use a computational framework which first extracts a rich set of nonverbal features (speaking activity, prosodic features, visual features like head nods, facial expressions using state of the art techniques) and verbal features like Linguistic Inquiry and Word Count (LIWC) and the state-of-the-art Doc2Vec features, and then uses machine learning methods for inference in regression tasks. Based on this framework, the contributions of this chapter are:

1. With respect to **RQ1**, we first conduct a cross-situation correlation analysis between perceived hirability and soft skills at job interviews, and perceived performance and soft skills at the reception desk. We find Pearson's correlation r in the range $[0.3, 0.49]$ implying that perceived variables in job interviews are moderate indicators of perceived performance and soft skills on the job. Second, we assess the inference of perceived performance and soft skills at the job using perceived variables in the interview setting, achieving a regression performance of $R^2 = 0.25$. The best performance ($R^2 = 0.40$) is achieved by fusing the perceived hirability and soft skills scores at the job interviews and nonverbal behavioral cues from the reception desk.
2. With respect to **RQ2**, we first conduct a Person's correlation analysis and observe that for both interview and reception desk, specific behavioral cues (longer speaking turn and head nods) are correlated to higher ratings of all perceived variables in the corresponding situation (r in the range $[-0.43, 0.39]$). We then conduct an inference experiment to infer perceived performance and soft skills using nonverbal behavioral cues from interviews. The best performance of $R^2 = 0.30$ is obtain by fusing nonverbal cues extracted from interview and desk situations.
3. With respect to **RQ3**, we conduct an inference experiment using linguistic content as input. We observe that the performance is lower than nonverbal behavioral cues, with best performance of $R^2 = 0.25$ using linguistic content features from reception desk setting only.

Our results have broader implications for human resources and managers in hospitality, by providing insights about potential employees' nonverbal behavior and its connections to

perceived performance on the job. Our work also contributes towards building a behavioral training program across situations with a focus on hospitality students. The material of the chapter was originally published in [107]

6.1 Data Corpus

We used a data corpus consisting of 169 interactions each in two situations; job interview and reception desk, described in Section 3.2. Here we again briefly outline the process of data collection and annotation of perceived variables.

The corpus consists of 100 students from the hospitality school who took part voluntarily. 69 students participated in the second session, while 31 did not return. The mean age of participants was 20.6 years, with 57 females and 43 males. The interactions were in either English or French (based on the choice of each participant) due to the international nature of the school, resulting in 260 (resp. 78) interactions in French (resp. English). Overall, the job interview dataset is 1690 minutes long (mean duration: 10 mins), while the reception desk dataset is 1350 minutes long (mean duration: 8 mins). In our investigation, we use the entire 338 videos (169 videos from each setting) and analyze at video-level. The two lab sessions were recorded 4 – 6 weeks apart. So we treat them as independent videos in line with ubicomp literature [72].

Both lab sessions were captured with multiple modalities. The video data of the interactions was recorded using two Kinect v2 devices (one for each interaction partner), and was recorded at 30 fps in RGB and depth (1920×1080 and 512×424 for RGB and depth, respectively.) Audio data was captured at 48kHz with a microphone array device that segmented speaker turns from localized sources. Audio and video streams are synchronized.

6.1.1 Annotations

The data was augmented with a number of manually labeled variables as described in [108]. The job interview videos were annotated by a group of five independent annotators, while the reception desk videos were annotated by a different group of three independent raters. Both groups of annotators were students and paid 20 CHF per hour for their work. The annotators in both groups rated the videos on various perceived variables on a seven-point Likert scale after watching the first two minutes of the videos (self-presentation in the job interview and complaint segment in the reception desk).

The perceived variables annotated for both situations along with their descriptive statistics are listed in Table 6.1. We use Intraclass Correlation Coefficient (ICC) [147], to measure the agreement between raters. Specifically, $ICC(2,k)$ is used as the measure of the inter-rater agreement because a sample of annotators was used, and each annotator judged all videos. From Table 6.1, we observe that the agreement among raters for all perceived variables was

Table 6.1 – List of perceived variables manually annotated for both situations, along with their $ICC(2,k)$ and means.

	Job Interview		Reception Desk	
Variable	$ICC(2,k)$	**Mean**	$ICC(2,k)$	**Mean**
Professional Skills				
Competent (compe)	0.56	6.01	0.69	4.24
Motivated (motiv)	0.52	5.89	0.63	4.80
Social Skills				
Positive (posit)	0.60	5.70	0.60	4.34
Sociable (socia)	0.57	5.67	0.64	4.46
Communication Skills				
Clear (clear)	0.67	5.89	0.66	4.56
Persuasive (persu)	0.69	5.57	0.72	4.01
Overall				
Performance (peImp)	–	–	0.77	4.11
Hirability (hire)	0.69	5.54	–	–

moderate to high with $ICC(2,k)$ in the range $[0.52, 0.77]$ for interview videos, while for the reception desk the $ICC(2,k)$ is in the range $[0.60, 0.77]$. ICC values greater than 0.5 are generally considered to be acceptable inter-rater agreement. For both situations, the distribution of all the perceived variables are centered on the positive side of the Likert scales (Mean ≥ 4) implying that both groups of annotators generally perceived the participants positively.

6.1.2 Speech Transcripts

To investigate the impact of linguistic content employed by participants in each situation, we used manually transcribed text from the audio tracks. The transcription was done by a pool of five master's students in organizational psychology, who were native French speakers and fluent in English, watched all the videos, and transcribed the interaction in the original language. The transcribed documents contained verbal content of both the research assistants' and the participants' speech. In our analysis, we use only the participants' data for two reasons: our focus is on participants behavior, and the research assistants' questions did not vary during the job interview situation.

6.2 Nonverbal and Verbal Feature Extraction

We extracted various nonverbal features from both job interview and reception desk videos. For extracting verbal features, we utilize the manually transcribed text data (outlined in Section 3.4). The complete description of the verbal and nonverbal features is presented in Section 3.5, we list them here briefly for sake of completeness.

1. **Nonverbal Features**

   a **Audio Features** include speaking activity features (composed of speaking time (total

time an individual speaks), speaking turns (segments greater than two seconds), pauses (gaps in speech shorter than two seconds), short utterances (speaking segments less than two seconds)) and prosody features (like pitch (voice fundamental frequency), speaking rate (speed at which words are spoken), spectral entropy (measure of irregularity or complexity), energy (voice loudness), voicing rate (number of voiced segments per second), and time derivative of energy (voice loudness modulation)).

b **Visual Features** include (a) Overall visual motion (b) head nods (c) visual back-channeling. These cues were captured and various statistics like count, mean, median, standard deviation, minimum, and maximum of duration were computed as features.

c **Multimodal Cues** are defined as events when protagonists nod their head while speaking. Count of nodding while speaking, mean, median, standard deviation, minimum, and maximum of duration were computed for use as features.

2. **Verbal Features**

   a **Linguistic Inquiry and Word Count**(LIWC) is a software [127] we use to extract lexical features.It computes these features by looking up each word in the transcript to the in-built English dictionary and is maps it to one of 70 categories.

   b **Doc2Vec** or paragraph vector was proposed by Le et al.[89] to represent documents.

6.3 Inference Framework and Experimental Protocol

In this section, we outline the inference framework and experimental protocol. The various input components for our experiments and their source are visualized in Figure 6.1. As a first step towards answering RQ1 and RQ2, we perform a correlation analysis between the variables of interest. We report only the correlation values which are significant with $p < 0.05$. We then define a regression task in which perceived performance and soft skills at the reception desk is inferred from nonverbal cues, ratings from job interviews, nonverbal cues from reception desk, linguistic content from both situations, and various combinations of these features.

For the regression tasks, we follow a standard machine learning protocol. First, the data was pre-processed using a person-independent Z-score normalization to transform data into unity variance and zero mean. Then, two unsupervised dimensionality reduction techniques were evaluated:

1. Low p-value features(p-val): In this method, features which were significantly correlated ($p < 0.05$) only were selected. This is based on the assumption that important information is encoded in significantly correlated features.
2. Principal Component Analysis (PCA): This method projects the features into a lower dimension orthogonal space [125].

The performance of these dimensionality reduction did not improve performance over the use of the original features and hence, their results are not reported here. Two regression

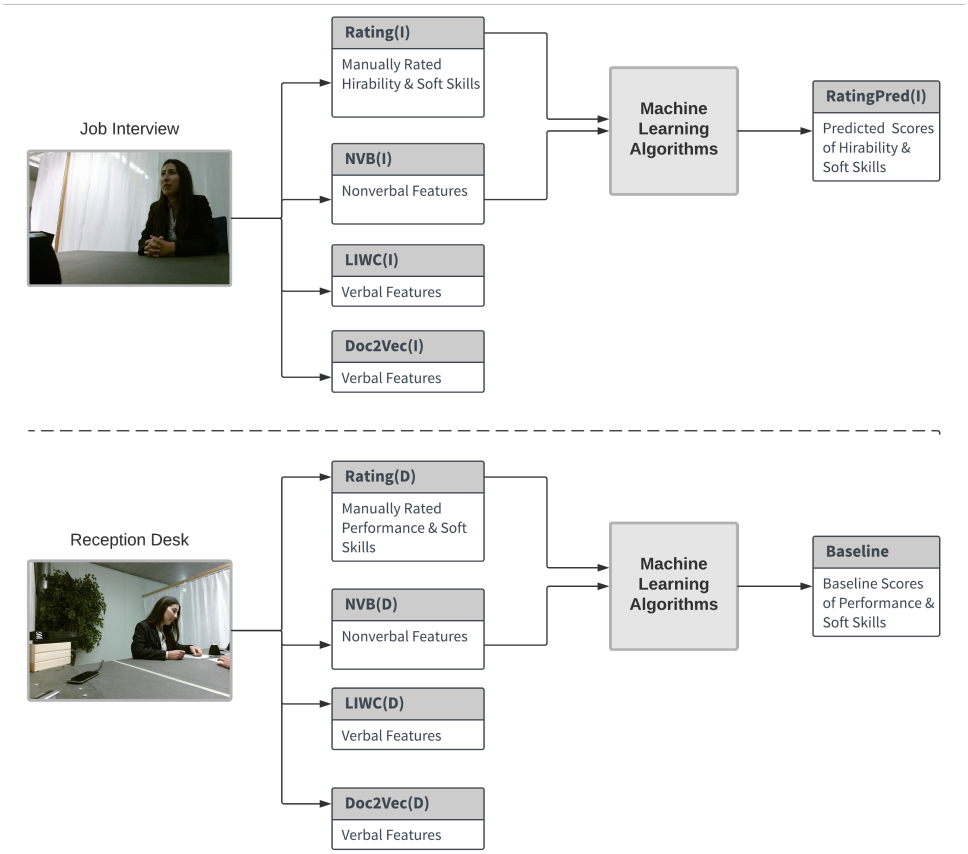


Figure 6.1 – A visual summary of the cues used in our experiments and how they were obtained.

techniques (Support Vector Machines regression (SVM-R) [35] and Random Forest regression (RF) [25]) were evaluated using the Caret package [88] for R implementation. These algorithms were selected to understand the contributions of each component in inferring perceived performance and soft skills in the reception desk situation. The hyper parameters of the machine learning algorithms were optimized using 10-fold inner cross-validation (CV), while the performance was assessed using the 100 independent runs of leave-one-video-out CV. The performance of these regression techniques was evaluated by employing coefficient of determination (R^2). We use the R^2 values reported in our previous work [110], obtained using nonverbal behavioral cues only, as the baseline for comparing results.

6.4 Results and Discussion

We now present the results and discussion corresponding to each of the three RQs we originally posed.

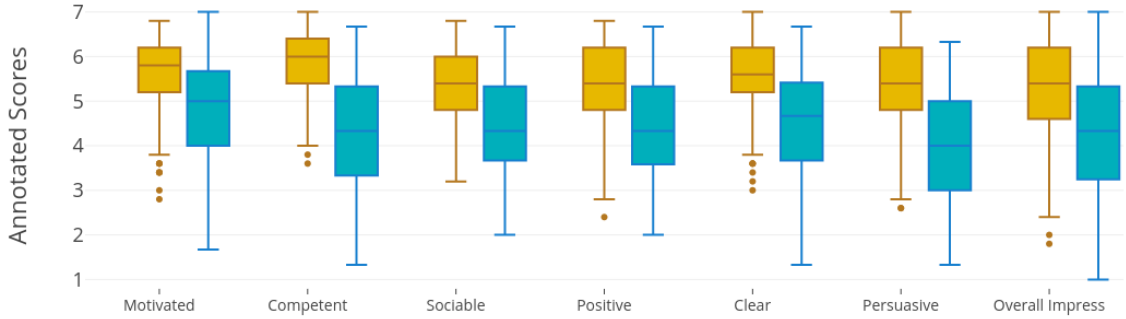


Figure 6.2 – Box-plot showing the distribution of annotated scores for each of the variable of interest. Here we observe that mean scores for interview (yellow) is greater than mean scores for reception desk (blue).

6.4.1 RQ1: Perceived Variables in Interview and Reception Desk Situations

We begin by computing the descriptive statistics of the perceived variables of both situations, presented as box plots in Figure 6.2. We observe that the mean ratings for all perceived variables in the reception desk situation are lower than the corresponding ratings in the job interviews, indicating that all variables were more favorably perceived in the interview than at the reception desk. We hypothesize that this is due to the reception desk interactions occurring in a more challenging situation (the client is unhappy and not easy to persuade) while the interview interactions occur under positive tone. This hypothesis has backing in psychology, which suggests that positive evaluations tend to occur under positive mood [96, 61]. To verify if the difference in perceived variables across the settings was significant, we conducted a test of means for each variable. As the population of participants was the same across both the settings, we used a paired Student T-test ($N = 169$). The test of means refuted the *Null* hypothesis ($p < 0.001$), indicating that the differences in mean perceived variable scores during the job interview were significantly higher than the mean perceived variable scores in the desk situation.

Correlation Analysis: We conducted a Pearson's correlation analysis on the perceived variables from the two situations. Results of this analysis are presented in Table 6.2. We observe

Table 6.2 – Pearson's correlation between perceived variables from interview (I) and reception desk (D) situations ($N = 169$). All of them are significant with $p < 0.001$

I.Motivated	I.Competent	I.Positive	I.Sociable	I.Clear	I.Persuasive	I.Hirability
0.49	0.41	0.44	0.49	0.30	0.40	0.45
D.Motivated	D.Competent	D.Positive	D.Sociable	D.Clear	D.Persuasive	D.Performance

Table 6.3 – List of predictors used in regression experiments obtained from job interview (I) and reception desk (D) interactions.

Abbreviation	Details
NVB (I)	Nonverbal behavior extracted from interviews
Ratings (I)	Manually rated hirability and soft skills from interviews
RatingPred(I)	Automatically predicted scores of hirability and soft skills from interviews
LIWC(I)	LIWC features extracted from manual transcriptions of interviews
Doc2Vec(I)	Doc2Vec features extracted from manual transcriptions of interviews
NVB (D)	Nonverbal behavior extracted from desk
Ratings (D)	Manually rated performance and soft skills from desk
LIWC(D)	LIWC features extracted from manual transcriptions of desk
Doc2Vec(D)	Doc2Vec features extracted from manual transcriptions of desk
LIWC(D + I)	Combined LIWC features extracted from manual transcriptions (interviews & desk)

that all perceived variables are positively correlated to each other ($p < 0.001$ in all cases). *Sociable* and *Motivated* in the two situations have the highest correlation ($r = 0.49$), while *Clear* has the lowest ($r = 0.30$). An interesting observation is the correlation between perceived performance and perceived hirability ($r = 0.45$). This seems to suggest that participants who were perceived as more hirable during their interviews were to some degree perceived to perform better on the job.

Inference Task: We then investigated the ability of the perceived variables from job interviews in inferring perceived performance and soft skills at the job situation as a regression task. The baseline for this work is the R^2 obtained using nonverbal cues to infer perceived variables, specifically *Performance* ($R^2 = 0.30$) reported in our previous work [110].

Table 6.3 summarizes the various predictors used in all our inference experiments. Towards answering **RQ1**, we define four experiments, labeled Exp1a-Exp1d Table 6.4, to test different conditions involving perceived variable scores (visualized in Figure 6.1). In *Exp1a*, we use the perceived scores from job interviews as predictors of perceived performance and soft skills. We observe that the best performance of these perceived variables (using SVM-R) was slightly lower than *Baseline* with $R^2 \in [0.18, 0.27]$. The best performance was observed for *Sociable* ($R^2 = 0.27$) and lowest for *Clear* ($R^2 = 0.18$). This set of predictors produces $R^2 = 0.25$ for *Performance*. These results can be explained by the correlations between the perceived variable in the two situations. Our results show that perceived performance and soft skills on the job can be inferred to some extent by just the perceived hirability and soft skills scores during interviews.

To further understand this connection, we conducted another regression task using automatically predicted scores from the job interviews instead of manually generated scores, to study a situation where fully automatic assessment at the interview could be used to make inference at the job (*Exp1b*). The predicted scores (RatingPred(I)) were obtained by using nonverbal cues displayed during job interviews as predictors in a regression task with random forest

Table 6.4 – Summary of experiments and the best regression performance (R^2) achieved. All results are significant with $p < 0.05$.

Experiment	Predictors	Model	motiv	compe	posit	socia	clear	persu	perfo
Baseline	NVB (D)	RF	0.30	0.29	0.32	0.33	0.22	0.32	0.30
Exp1a	Ratings (I)	SVM-R	0.21	0.23	0.26	0.27	0.18	0.24	0.25
Exp1b	RatingPred(I)	SVM-R	0.27	0.26	0.28	0.26	0.16	0.22	0.21
Exp1c	Ratings (I) + NVB (D)	RF	0.34	0.37	0.36	0.39	0.24	0.37	0.40
Exp1d	RatingPred(I) + NVB(D)	RF	0.36	0.32	0.32	0.34	0.23	0.30	0.31

(RF). This method has been shown to result in $R^2 = 0.32$ [108, 114]. A paired test of means accepted the null hypothesis, indicating that the means of predicted scores and manual scores were not statistically significant. However, the use of these predicted scores (RatingPred(I)) for regression at the job showed a lower performance for four of the variables like *Clear, Persuasive* and *Performance* ($R^2 \in [0.16, 0.28]$). Even though the performance was lower, this result shows a first step towards using automatically inferred scores of job interviews to infer perceived performance and soft skills at the job.

The best performing model was obtained in *Exp1c*, where we studied the effect of combining nonverbal cues displayed at the desk and scores of perceived variables from interviews (using RF). We obtain $R^2 = 0.40$ for *Performance*, compared to a baseline of $R^2 = 0.30$. An improved inference performance is also observed for other variables with *Sociable* ($R^2 = 0.39$), *Competent, Persuasive* ($R^2 = 0.37$), *Positive* ($R^2 = 0.36$), with the lowest performance for *Clear* ($R^2 = 0.24$). To complete the experiments, we infer the impressions at desk using automatically predicted scores from the interview (*Exp1d*) in addition to nonverbal cues extracted from reception desk. The results indicate that this fully automated condition brings about marginal improvement over the baseline like *Performance* (from 0.30 to 0.31), *Clear* (from 0.22 to 0.23), *Sociable* (from 0.33 to 0.34), and *Competent* (from 0.29 to 0.32).

As the next step, to understand the contributions of features to infer perceived performance, we list the top 20 variables used by the RF algorithm (Table 6.5). This list was obtained by using the *var.Imp* function in CARET, which returns the variables and their measure of importance (scaled to 100). We observe that this list of top variables includes scores of perceived variables from job interviews, nonverbal cues from both the participants and the clients. Specifically, we observed that *Hirability* and *Persuasive* scores rated at the interview were marked as two of the seven most important variables by RF. Similarly, participant cues found to contribute include speaking time, turn duration (mean and max), head nods (mean and duration), voice energy modulation (upper and lower quartile), and visual back-channeling (duration). An interesting observation is that client nonverbal cues like speaking energy, voice energy modulation, and spectral entropy also contribute to inference performance.

To summarize, in this subsection we investigated the question: How are perceived variables in the job interview connected with perceived variables on the job situation? Our main results are: (1) Scores of perceived variables from job interviews and perceived variables in reception desk are moderately correlated. (2) The perceived variables scores at reception desk can be inferred to some extent ($R^2 \in [0.21, 0.25]$) from perceived variables in job interviews, both

Table 6.5 – Top 20 variable importance in the RF for *Exp1c*. All measures of importance indicated in the *Rank* column are scaled to have a maximum value of 100.

Cues	Rank	Cues	Rank
Participant cues			
Speaking time	100.00	Total number of head nods	64.41
Speaking ratio	80.01	Upper quartile of change in speaking energy	62.98
Mean duration of nodding while speaking	74.25	Std of turn duration	62.22
Number of nods while speaking	73.22	Lower quartile of change in speaking energy	62.08
Mean turn duration	72.83	Max duration visual back-channeling	59.18
Max turn duration	67.15		
Interview Ratings			
Persuasive	87.67	Motivated	58.44
Hirability rating (Interview)	69.56	Communicative rating (Interview)	57.18
Enthusiastic	61.05		
Client cues			
Lower quartile speaking energy	67.78	Min spectral entropy	59.45
Upper quartile of change in speaking energy	60.15	Max speaking energy	57.18

manual rated and automatically inferred. (3) The fusion of the perceived variable scores from job interview and nonverbal cues extracted from the desk improves inference of perceived variables at the desk, with a best performance of $R^2 = 0.40$. Our results indicate that the impressions made during job interviews add information to the nonverbal behavior during the desk situation.

6.4.2 RQ2: NVB in Interviews and Perceived Performance at the Desk

In this subsection, we investigate the links between automatically extracted nonverbal behavior of candidates during job interviews and the perception of performance on the job. We first present a correlation analysis and then the inference task.

Correlation Analysis: As a first step, we conduct a Pearson's correlation analysis between nonverbal cues extracted during the job interviews and perceived variables at the desk situation. The results that show weak to moderate trends are presented in Table 6.6. We observe that participants who spoke for longer duration, with less silence, and had greater speaking energy modulation during the job interview, were perceived to perform better at the reception desk. Also, participants who nodded more, for greater duration, displayed greater number of visual back-channeling, and nodded more while speaking were rated as better performing during the reception desk.

These results are supported by literature in psychology [40, 104]. In [104], Motowidlo et al. using a dataset of simulated job interviews of 40 managers reported correlations of $r = 0.32$ between visual features and performance ratings, $r = 0.33$ between aural features and performance ratings, and $r = 0.36$ between combined aural and visual features and performance ratings. In that work, supervisors' ratings were considered as performance ratings. Similar results were reported in another work by DeGroot et al. [40]. Using videotaped interviews of 110 managers in a news-publishing company, it was reported that vocal cues correlated with performance ratings with $r = 0.20$ ($p < 0.05$). That work also found low correlations of $r = 0.14$ ($p < 0.05$) between performance ratings and composite visual cues (like physical

Table 6.6 – Pearson's correlation between perceived variables of desk and nonverbal cues displayed during job interviews ($N = 169$). All features are significant $^{***}p < 0.001$,$^{**}p < 0.01$, $^{*}p < 0.05$

NVB(I)	Ratings(D)						
	motiv	compe	posit	socia	clear	persu	peImp
Speaking Activity Features							
Num speaking turns	-0.23**	-0.24**	-0.18*	-0.15	-0.26**	-0.28**	-0.21*
Mean turn duration	0.33***	0.24**	0.21*	0.17*	0.23**	0.25**	0.24**
Number of silence events	-0.36***	-0.30***	-0.31***	-0.32***	-0.27**	-0.35***	-0.30***
Silence Ratio	-0.36***	-0.31***	-0.31***	-0.33***	-0.29***	-0.34***	-0.27**
Prosodic Features							
Lower quartile speaking energy	0.27**	0.22**	0.31***	0.35***	0.21*	0.22*	0.23**
Max speaking energy change	0.24**	0.22**	0.22*	0.28***	0.21*	0.18*	0.22*
Visual Features							
Total num head nod	0.41***	0.36***	0.38***	0.45***	0.33***	0.32***	0.33***
Total duration head nod	0.39***	0.38***	0.37***	0.43***	0.35***	0.36***	0.35***
Num of nod speak	0.37***	0.29***	0.32***	0.39***	0.26**	0.25**	0.28**
Mean duration of nod speak	0.35***	0.28**	0.22*	0.31***	0.23**	0.26**	0.27**
Std duration nod speak	0.34***	0.28***	0.24**	0.33***	0.22**	0.26**	0.26**
Max duration nod speak	0.40***	0.33***	0.30***	0.40***	0.26**	0.30***	0.30***
Num visual BC	0.31***	0.33***	0.31***	0.36***	0.31***	0.32***	0.29***
Mean duration visual BC	0.23**	0.30***	0.27**	0.27**	0.31***	0.31***	0.28**
Std duration visual BC	0.23**	0.28**	0.25**	0.29***	0.26**	0.26**	0.27**
Max duration visual BC	0.28**	0.30***	0.28***	0.30***	0.28**	0.29***	0.29***

attractiveness, smiling, gaze, hand movement, and body orientation).

We then compute Pearson's correlation between nonverbal behavioral cues in the two situations and perceived variable scores (Table 6.7). Specifically, we compute (a) correlation between nonverbal cues extracted from interviews (NVB(I)) and perceived hirability (Rating(I)) (b) correlation between nonverbal cues extracted from reception desk (NVB(D)) and perceived performance (Rating(D)). We observe that participants who displayed specific behavioral patterns had a weak-to-moderate trend to be rated high in both situations. Specifically, participants who spoke longer, louder, had fewer silence events were perceived as more hirable during the interview and also perceived as better performing on the job. Similarly, participants were perceived more positively when they moved more, nodded more and for longer time. This suggests that positive impressions could be related to similar behavioral cues in the two situations. Our results are in accordance with existing literature for interviews, where participants speaking for longer, with fewer silence, nodded more were rated as more hirable [39, 114, 108].

Inference Task: We then use regression to infer the perceived performance and soft skills at the reception desk from automatically extracted nonverbal cues from the interviews. A total of four experiments labeled Exp2a-Exp2d, were conducted using the various components illustrated in Figure 6.1 and the results are presented in Table 6.8. As a first step (*Exp2a*), we use all the nonverbal cues displayed during interviews to infer perceived performance and soft skills at the desk. We observe that these nonverbal cues overall have low predictive power with $R^2 \in [0.12, 0.30]$. The best performance is achieved for *Sociable* ($R^2 = 0.30$) and lowest for *Clear* ($R^2 = 0.12$). For *Performance*, this model achieved $R^2 = 0.17$. Though these results are lower then the baseline, they indicate a weak connection between behavioral cues from job

Table 6.7 – Selected Pearson's correlation coefficient for perceived hirability (Rating(I)) and perceived performance (Rating(D)) across the two situations ($N = 169$). $^{**}\,p < 0.01$, $^{*}\,p < 0.05$

Nonverbal Cues	Perceived Hirability	Perceived Performance
Acoustic Features		
Avg Turn duration	0.39^{**}	0.40^{**}
Speaking Ratio	0.21^{**}	0.43^{**}
Num Silent Events	-0.43^{**}	-0.22^{**}
Speaking Energy (Q25)	0.29^{**}	0.18^{*}
Speaking Energy Derivative (Q25)	-0.27^{**}	-0.20^{*}
Visual Features		
Mean WMEI	0.18^{*}	0.26^{*}
Max WMEI	0.16^{*}	0.30^{*}
Total Head Nod	0.25^{**}	0.37^{**}
Num of Nods while speaking	0.26^{**}	0.45^{**}
Max duration of Nods while speaking	0.25^{**}	0.39^{**}

interview and perceived variables on the job.

In the next step, we fuse the nonverbal cues extracted from the two situations and use them as predictors. The idea is to investigate the effect of extra behavioral information on the inference performance. In *Exp2b*, we observed that inference of some variables improved as compared to the baseline. Specifically, there is improvement for *Motivated* (from 0.30 to 0.34), *Sociable* (from 0.33 to 0.36) and *Competent* (from 0.29 to 0.30) while for *Positive* and *Persuasive* the performance decreased slightly. Fusion of nonverbal cues from both situations had no effect on inference of *Performance* at desk.

We then combine nonverbal cues and perceived scores from interviews to infer perceived variables at reception desk (*Exp2c*). The performance varies with $R^2 \in [0.26, 0.35]$ with best performances for *Positive* and *Sociable* ($R^2 = 0.35$) followed by *Motivated* ($R^2 = 0.32$), *Competent, Persuasive* ($R^2 = 0.28$) and *Performance* ($R^2 = 0.26$). This is the best result achieved using all information available from the job interviews and is comparable to *Baseline*, importantly, without seeing *any* data at the job. As a final experiment (*Exp2d*), we fused perceived scores at job interviews and all the nonverbal cues extracted from both situations and use them as predictors. We observe a slightly improved performance compared to the *Baseline* with the highest variance explained for *Sociable* ($R^2 = 0.39$), followed by *Positive, Persuasive*, and *Performance* ($R^2 = 0.32$).

To summarize, there are two main findings in this subsection: (1) Some nonverbal cues like speaking and turn duration, head nods displayed during job interviews are weakly-to-moderately correlated to perceived performance and soft skills at the reception desk. This result could have implications for behavioral training systems where focus can be on specific behaviors for multiple situations. (2) We observed that nonverbal cues extracted from job

Table 6.8 – Summary of experiments and best regression performance (R^2) of desk perceived variables achieved. All results are significant with $p < 0.05$.

Experiment	Predictors	Best Model	motiv	compe	posit	socia	clear	persu	perfo
Baseline	NVB (D)	RF	0.30	0.29	0.32	0.33	0.22	0.32	0.30
Exp2a	NVB (I)	RF	0.24	0.18	0.17	0.30	0.16	0.12	0.17
Exp2b	NVB (I) + NVB (D)	RF	0.34	0.30	0.27	0.36	0.22	0.29	0.30
Exp2c	NVB (I) + Ratings (I)	SVM-R	0.32	0.28	0.35	0.35	0.25	0.29	0.26
Exp2d	NVB (I) + Ratings (I) NVB (D)	RF	0.33	0.28	0.32	0.39	0.27	0.32	0.32

interviews have weak inference ability ($R^2 = 0.17$). Importantly, this performance improves ($R^2 = 0.26$) when these nonverbal features are augmented with perceived scores from job interviews. These results suggest that for some soft skills displayed in the actual job, it is useful to use behavior and impressions from the interview situation.

6.4.3 RQ3: Linguistic Content and Perceived Performance

To address *RQ3*, we conducted nine experiments Exp3a-Exp3i, with different linguistic features extracted. Here again the *Baseline* is the performance obtained in inferring perceived performance using nonverbal cues extracted from the reception desk ($R^2 = 0.30$). The input for these experiments is illustrated in Figure 6.1 and the results are tabulated in Table 6.9.

LIWC: First, we use LIWC to extract lexical cues from the reception desk transcribed data (LIWC(D)) and use them to infer perceived performance and soft skills (*Exp3a*). LIWC features show lower performance than the *Baseline* with $R^2 \in [0.09, 0.24]$ for all variables. The best performance was for *Clear* ($R^2 = 0.24$) followed by *Competent* ($R^2 = 0.22$), and *Motivated* ($R^2 = 0.09$) being the worst. Linguistic content of the desk results in $R^2 = 0.18$ for *Performance*. Note that this is better than results reported in investigations of linguistic content and *Overall Impression* in job interviews ($R^2 = 0.11$) in the literature [112, 106].

In a second step (*Exp3b*), using LIWC features extracted from job interviews (LIWC(I)), we find that the performance of linguistic content in inferring perceived performance and soft skills is very low, with $R^2 < 0.1$ for almost all variables (except *Clear*, $R^2 = 0.13$). We then combine the LIWC features from both settings (LIWC(D+I)) to infer impressions of performance and skills (*Exp3c*). We find no improvement except for *Motivated*.

Doc2Vec: We then investigate the potential of Doc2Vec with features extracted using the reception desk, Doc2Vec(D) in *Exp3d*. Interestingly, the performance of the Doc2Vec(D) is lower than the LIWC(D) features with $R^2 \in [0.05, 0.10]$, with $R^2 = 0.10$ for *Performance*. This is in contrast to results reported in the literature for job interviews [28]. In that work, the authors using Doc2Vec features to infer *Hirabilty* scores from 36 job interviews, and reported a correlation $r = 0.41$ between manual and automatic hirability. Converting r to R^2 for comparison, this work achieved $R^2 = 0.16$. We believe that the low performance we obtain

Table 6.9 – Summary of experiments with linguistic content and the best inference performance achieved. All results are significant with $p < 0.05$.

Experiment	Predictors	Model	motiv	compe	posit	socia	clear	persu	perfo
Baseline	NVB (D)	RF	0.30	0.29	0.32	0.33	0.22	0.32	0.30
Exp3a	LIWC(D)	RF	0.09	0.22	0.19	0.14	0.24	0.17	0.18
Exp3b	LIWC(I)	RF	0.04	0.04	0.02	0.01	0.13	0.02	0.07
Exp3c	LIWC(D + I)	RF	0.15	0.22	0.19	0.14	0.25	0.17	0.18
Exp3d	Doc2Vec(D)	SVM-R	0.08	0.09	0.07	0.05	0.06	0.08	0.10
Exp3e	Doc2Vec(I)	SVM-R	0.16	0.18	0.18	0.15	0.10	0.09	0.16
Exp3f	LIWC(D) + Doc2Vec(D)	SVM-R	0.17	0.22	0.16	0.19	0.19	0.20	0.25
Exp3g	LIWC(I) + Doc2Vec(I)	SVM-R	0.24	0.24	0.17	0.17	0.18	0.16	0.26
Exp3h	LIWC(D) + NVB(D) Doc2Vec(D)	RF	0.25	0.29	0.27	0.32	0.27	0.29	0.26
Exp3i	LIWC(I) + NVB(I) Doc2Vec(I)	RF	0.28	0.18	0.16	0.29	0.11	0.15	0.20

could be due to the relatively short duration of the reception desk interactions, which has an average of 354.1 words for all turns taken by the participant. The authors of [28] have not reported the corpus size used in their work so a direct comparison is not possible. As a next step, we use the Doc2Vec(I) features consisting of Doc2Vec features extracted from the job interviews (*Exp3e*). Interestingly, the performance with this feature set was better than the one achieved by Doc2Vec(D), with $R^2 \in [0.09, 0.18]$. The use of these features produces $R^2 = 0.16$ for *Performance* and is similar in range to those reported by Chen et al. [28]. We believe this improvement in performance might be due to the larger duration of the job interviews. This corpus is more than twice as long as the reception desk corpus, and contains an average of 813 words.

Fusion of LIWC & Doc2Vec: As a next step, we combine the two linguistic features. In *Exp3f*, we combine LIWC(D) and Doc2Vec(D) as predictors. We observe that the inference performance is better than each of the features individually, with $R^2 = 0.25$ for *Performance* as compared to $R^2 = 0.10$ and $R^2 = 0.18$ for Doc2Vec(D) and LIWC(D), respectively. The improvement is observed for all the perceived variables of the reception desk. Similarly, the fusion of linguistic features from the interviews (*Exp3g*) also leads to an improved performance for all the perceived variables as compare to each of individual features. With the fused feature set, we observe $R^2 = 0.26$ for *Performance* is explained. This is the best performance achieved using the linguistic features.

Fusion of Linguistic and Nonverbal Features: In the final step, we use a fusion of nonverbal cues and linguistic features from the reception desk situation as predictors (*Exp3h*). Except for one variable (*Clear*, $R^2 = 0.27$) these results are not better than the *Baseline* performance. The same is the case when we combine nonverbal and linguistic features from the interview situation (*Exp3i*), which do not improve over the baseline (NVB(D)). To understand this result, we listed the top 20 variables used by the RF algorithm using the *var.Imp* function in CARET. We do not report them here as this list did not contain any verbal features in top 20 and hence

Questionnaire for hospitality students:

1. How does the role playing during the interview experiment relate to your real-world experience?
2. How does the role playing during the front-desk experiment relate to your real-world experience?
3. Did you know that there are connections between how well you did in the interview and how well you do on the job.
4. How in your opinion can technology help you improve your behavior during job interviews?
5. How in your opinion can technology help you improve your behavior on the job?

Figure 6.3 – List of questions sent to hospitality students as a part of a qualitative study to understand the implication of this work.

not very helpful in understanding the impact of linguistic features on perceived variables.

To summarize, the main findings of this section are (1) LIWC features outperforms the Doc2Vec features using the reception desk data, with the best performance being always worse than using nonverbal cues. (2) Interestingly, the Doc2Vec features from job interviews perform comparable to LIWC features from desk. (3) The fusion of LIWC and Doc2Vec features from both situation results in improved inference, with *Exp3f* giving $R^2 = 0.25$, while *Exp3g* gives $R^2 = 0.26$. Overall, linguistic features can, moderately and to a lesser degree than nonverbal behavior, be useful to infer perceived performance and soft skills.

6.4.4 Qualitative Study

To understand the implications of this work for real-world situations in hospitality, we conducted a small qualitative study. The study consisted of two sets of questionnaires consisting of five questions (Figure **??** and Figure **??**). One set of questions was sent to ten selected participants of the study, while the other was sent to two directors of the hospitality school where the dataset was collected. Of the 12 people contacted, we received responses from four people, two participants (henceforth called student A and B) and two directors (henceforth called director C and D).

Specifically, we asked the hospitality student recipients of the questionnaire about their experience during the interview role play and its relation to the real world (Figure **??**). Student A said that “It was a good experience to realize how stressful an interview could be. I was very happy to do it because a few months after I had to do a ‘true’ interview and they asked me similar questions! I was feeling prepared because I knew how to deal with it. Like if I prepared an exam.” Student B, replying to a question about the experience during the reception desk role play and its relation to the real world, said “It is a very common situation that is faced in reception so it was a very appropriate exercise linked to our line of work. Being able to handle

Questionnaire for hospitality directors:

1. In your experience, how do you use the information from job interviews to forecast on-the-job performance of a young hospitality employee?
2. How, in your opinion, can technology help young hospitality professionals to improve their behavior during job interviews?
3. How, in your opinion, can technology help young hospitality professionals to improve their behavior on the job (like the front desk)?
4. What is the value of role-playing job interviews in relation to the real-world experience of job interviews for young hospitality professionals?
5. What is the value of role playing in relation to the real-world experience of front desk interactions for young hospitality professionals?

Figure 6.4 – List of questions sent to hospitality school directos as a part of a qualitative study to understand the implication of this work.

dissatisfied guests or situations under pressure is good practice for us to learn, how to foresee situations or be proactive within our line of work for the future". Both students said they did not know that there might exist a link between interview ratings and perceived performance on the job. Similarly, both students felt that the use of technology in specific stressful situations (like angry clients at a reception desk) can help them improve their nonverbal behavior. Specifically, student A said, "It can give us a perspective that we may not have noticed before or change our opinion on a certain behavior".

The two directors of the hospitality school too felt that the role-playing during data collection had connections to the real world situations in hospitality (questionnaire in Figure **??**). Specifically, director C expressed the opinion that role-playing could help young hospitality professionals as "One can play out different scenarios about guest contact, without actually throwing the person in at the deep end. In other words, one can practice without real guests' difficult situations which in turn will assist the young professional once he/she encounters them." Director D was on the opinion that role-playing helps students gain insights on the nuances of what the job entails. He, however, cautioned that "in the real world, even if our students have the required skills and personality to be hired in a position, the most challenging issues will be to adapt to a new people and environments as well as to adopt new style of work."

Furthermore, the two directors were enthusiastic about the role of technology in improving young professionals' behavior on the job. Director C responded by saying that "technology can be used as a mirror of actions and to effectively communicate desired behavior patterns that can serve as a role-model for people on the job". Director D felt that by using technology "students and young professionals could train and improve their behaviors and speeches when

faced with different types of clients: introverts, extroverts, violent, sly".

In summary, both the participants and the directors of the school felt that role-playing of job-related situations helped hospitality students be prepared for stressful situations like facing an interview or a difficult client at the reception desk. They were also enthusiastic about using ubiquitous computing systems to capture and analyze behavioral cues as they felt they can help achieve better behavioral awareness during professional interactions in customer-facing jobs.

6.5 Conclusion

This paper describes our investigation into human behavior (verbal and nonverbal) and formation of impressions across multiple situations using ubiquitous sensing and multimodal analysis. Specifically, we investigated the connections between verbal content, displayed nonverbal behavior, and perceived variables under two different situations in the context of hospitality. Towards this, we used a data corpus of 338 interactions, recorded in multiple modalities and role-played by hospitality students in two settings; job interview and reception desk. A number of nonverbal behavioral cues were automatically extracted. Further, the interview and desk interactions were manually transcribed, and then verbal cues were extracted from these transcriptions.

We posed three research questions (RQs) summarized here:
RQ1 examined the connections between perceptions of candidates in job interviews and perceptions of the same person on the job. The four main findings were: (1) mean scores of perceived variables were higher in the job interviews than the corresponding ratings on the job. This implies that participants were perceived more favorable during the interviews. (2) We observed that perceived variables from job interviews were weakly to moderately positively, correlated to perceived variables in the reception desk situation. (3) We found that perceived variables on the job can be inferred, to some extent, from manually rated perceived variables ($R^2 = 0.25$) and automatically inferred scores ($R^2 = 0.21$) in job interviews. (4) The fusion of automatically extracted nonverbal cues from the desk situation with the perceived variable scores from interviews improved inference of perceived variables on the job, and corresponding best performance ($R^2 = 0.40$).

RQ2 examined the link between automatically extracted nonverbal behavior of candidates during job interviews and the perception of performance on the job. There were two main findings: (1) Some nonverbal cues displayed during job interviews were weakly to moderately correlated to perceived performance and soft skills at the reception desk situation. (2) Using these nonverbal cues as predictors in an inference task had a moderate performance with $R^2 = 0.17$ for perceived performance. Augmenting these nonverbal features with perceived variable scores in job interviews, the performance improved with $R^2 = 0.26$. Our result indicates a moderate feasibility to use nonverbal cues displayed during job interviews in inferring perceived performance and soft skills in the reception desk setting.

RQ3 studied the connections between candidates' choice of words in the two interactions and the perception of performance on the job. This results revealed some feasibility of using linguistic features to infer perceived variables on the job, although their performance is lower for all the perceived variables than the baseline. The three main findings were: (1) LIWC features extracted from reception desk outperformed the Doc2Vec features computed from the same situation in the inference of perceived variables on the job. (2) The Doc2Vec features extracted from job interviews performed comparably to LIWC features extracted from the reception desk situation. (3) Fusing the LIWC and Doc2Vec features from the desk situation improved inference performance, with $R^2 = 0.25$ achieved for Performance.

7 Other Applications

This chapter presents two applications of our findings thus far in other data sources. In this thesis so far, we have shown the feasibility of using verbal content to infer first impressions albeit with low inference performance using data recorded in lab environment. In Section 7.1, we investigate connections between the choice of words and first impressions in a data corpus consisting of noisy, "in-the-wild" video resumes from YouTube, using existing NLP representations. This is motivated by the fact that most of the existing research in work-related contexts has focused on nonverbal behavior with verbal content receiving little attention. In Section 7.2, we evaluate a real-time feedback system leveraging advances in wearable computing. Our research so far has shown the importance of speaking time in the formation of first impressions independent of a given situation. Using this result, a real-time feedback system was developed on an Android platform. This work was conducted in collaboration with Jean Costa from Cornell University.

7.1 Verbal Content and Hirability Impressions in Video Resumes

This thesis so far has investigated nonverbal and verbal behaviors and their connections to first impressions in two varied workplace situations. Specifically, the job interviews we explored previously were face-to-face interactions (i.e both the participants were sitting across a table in the same room during the interaction). Job interviews can be categorized into three types based on the interaction. First, *Face-to-face interviews* are the traditional method where interviewer and interviewee sit facing each other and have been widely investigated in social computing [114, 112, 108, 28]. Second, *video interviews* constitute a setting where participants answer questions in front of a computer, similar to face-to-face interviews, but without the presence of an interviewer. These have been investigated in the context of Big-5 personality [15], hirability [30], and online training systems [53, 72]. Third, *video resumes* are short videos in which job applicants present themselves and their communication skills to potential employers [68]. The wide-spread acceptance of social media like YouTube has led to the emergence of such videos. Although this new type of media allows researchers to study work-related social constructs and first impressions at large scale, video resumes yet remain

relatively seldom investigated from a behavioral standpoint [116].

Most of the existing research in work-related contexts has focused on visual and aural nonverbal behavior. Verbal content has received little attention as manual transcriptions of social interactions are a time consuming and expensive process. Research in psychology has shown that the words we use (in both written and spoken form) are influenced by various aspects of our identity [32]. Choice of words also provide insights into our thought processes, emotional states, intentions, and motivations [159]. Hence, verbal behavior also plays a role in how we are perceived by others; in this sense, analyzing verbal content is an important step in the understanding formation of first impressions.

In the context of face-to-face [106, 112, 28] and video interviews [30], previous works have investigated the relationship between verbal content and the formation of first impressions. To the best of our knowledge, no previous work has analyzed verbal content in video resumes. The closest to this work is from Biel et al. who investigated the relation between verbal content and personality impressions using video blogs (vlogs) [20] from YouTube but the relatively high word-error rate of the automatic speech transcription used at the time of the study, degraded the quality of inference. In this study, we investigate the role of verbal content in the formation of the first impression in the context of online conversational video resumes. To the best of our knowledge, we are the first to utilize advances in natural language processing (Doc2Vec, Word2Vec, GloVe) to understand verbal behavior in this context. Towards this, we use 292 YouTube video resumes and address the following research questions.

1. RQ-1: How can verbal content be represented for the inference of hirabililty impressions in video resumes?
2. RQ-2: What is the effect of automatic speech recognition on inference performance compared to manual transcription?
3. RQ-3: What is the impact of video duration on the inference performance of verbal content in hirability impressions?

Towards this goal, we develop a computational model to automatically extract various verbal representations from text corpus and evaluate their performance in a regression task. The contribution of this work are:

1. We transcribe the first 2 minutes of a randomly selected subset of 292 videos both manually and automatically using Google Speech API
2. We extract various representations of verbal content including LIWC, Doc2Vec, Word2Vec and GloVe
3. We evaluate the various representations in an inference task and observe that the highest inference performance is obtained for *Overall Hirability* ($R^2 = 0.23$) using the GloVe model.
4. We then assess the performance of automatic transcription versus manual and observe comparable inference performances, with $R^2 = 0.21$ for *Overall Hirability*.
5. We assess the difference in performance between automatic transcription of 2 minutes

versus full video duration and observe that inference performance improve slightly with $R^2 = 0.22$ for *Overall Hirability*.

The material in this section was originally published in [109].

7.1.1 Related Work

Nonverbal Behavior In Interview-Related Settings

In organizations, job interviews constitute among the most widely used tools for hiring the best applicant; for this reason, they have been widely studied in psychology and computing. Traditionally, psychologists have investigated job interviews from a nonverbal standpoint. It has been established that the applicant's nonverbal behavior influences the hiring decisions of the recruiter. Specifically, more eye contact, smiling (Imada et al. [77]), more facial expressions, nodding (Forbes et al. [48]), voice modulation, and fluent speech (McGovern at al. [98]) were shown to have a positive influence on the outcome of interviews.

Advances in ubiquitous sensors and improved perceptual techniques have enabled the automatic analysis of face-to-face job interviews. Nguyen et al. [114] automatically extracted a number of acoustic and visual nonverbal cues from a dataset of 62 real interviews to infer hirability and personality impressions, and reported a performance of $R^2 = 0.36$ for hirability variables. This work was extended by Naim et al. who investigated hirability and various other social constructs e.g. friendliness, engagement in a dataset consisting of 138 simulated job interviews. Using automatically extracted nonverbal features (including facial expressions, prosody), the authors reported correlation coefficients up to $r = 0.70$. Muralidhar et al. designed and developed a behavioral training procedure to help hospitality students improve their first impressions [108]. Using this live-in lab, they collected 169 simulated interviews and reported an inference performance of $R^2 = 0.32$ using nonverbal behavioral cues.

Related to face-to-face job interviews are *video interviews*, consist of participants answering questions in front of a computer without the presence of an interviewer. Batrinca et al. [15] were the first to analyze this type of setting; they investigated the formation of Big-Five personality impressions using a dataset of 89 participants . Their system automatically extracted 29 simple acoustic and visual nonverbal features and reported that Conscientiousness and Neuroticism were best recognized traits. Chen et al. investigated hirability and personality impressions using 1891 video interviews [30]. In a similar setting, using multimodal cues (i.e aural, visual and text), the authors reported F-measures ranging from 0.6 to 0.8 for personality and hirability impressions in a classification task.

The enormous popularity of social video platforms like YouTube has enabled the emergence of *video resumes*, which are short videos in which job-seekers present themselves and their communication skills to potential employers [68]. Although, these online videos open up new avenues for researchers to study work-related social constructs and first impressions at

large scale, the related work studying video resumes from a behavioral stand point is scare. To our knowledge, Nguyen et al. [116] is the only work focusing on this setting. Specifically, the authors collected 939 English-speaking conversational video resumes from YouTube. They automatically extracted acoustic (speaking activity, prosody) and visual (proximity, frontal face events, and head motion) nonverbal cues and analyzed their relationship with Big-Five personality and hirability variables. In a regression task, they reported an inference performance of $R^2 = 0.27$ for Extraversion, and up to $R^2 = 0.20$ for social and communication skills.

Verbal Behavioral In Interview-Related Settings

Research in social psychology have indicated the relation between choice of words, and our thought processes and emotional states [159] and thus verbal behavior plays an important role in impression formation. In the context of job interviews, literature in psychology has investigated the role of verbal behavior in the formation of impressions of job-related social constructs[71, 70]. Hollandsworth et al. analyzed 338 on-campus job interviews and reported that appropriateness of content contributed to favorable outcome of job interviews [71]. Rasmussen reported similar results using 80 simulated interviews of undergraduate students [134]. The author reported that positive interview outcomes were influenced by relevant verbal content along with consistent nonverbal behavior.

A common method to represent verbal content is using Linguistic Inquiry and Word Count (LIWC) software. LIWC has been extensively used to validate the psychometric properties of words [127]. This software was based on the existence of relationship between choice of words and persons' thoughts, emotional states and motivations [32, 159]. The representation is achieved by looking up each word in the text corpus to an in-built language dictionary and is mapped to one of 70 categories. LIWC has been used in computing literature too, investigating relationships between linguistic style and personality [20], leadership [143], and hirability impressions [28, 106, 112].

In the context of videos from social media, Biel et al. investigated the relation between verbal content and personality impressions in YouTube video blogs (vlogs) [20]. In this work, the authors used manual and automatic speech transcriptions to understand their impact. They used features extracted using LIWC software and reported an inference performance of $R^2 = 0.31$ for *Agreeableness* using the manual transcriptions. The use of automatically transcribed speech using two two-pass systems that use acoustic models for English, had a very poor performance ($R^2 = 0.18$), but this can be attributed to the high word-error rate of the ASR system (62.4%).

Table 7.1 – Inter-rater agreement ($ICC(1,k)$) and descriptive statistics for crowd-sourced annotations with each video being annotated by 5 raters. Total number of videos are 939 (Source: Nguyen et al. [116])

Variables	$ICC(1,k)$	Mean	STD
Overall Impression	0.59	3.70	0.62
Overall Hirability	0.61	3.72	0.62
Professional Skills	0.59	3.76	0.60
Social Skills	0.57	3.67	0.63
Communication Skills	0.64	3.71	0.69

7.1.2 Dataset

YouTube Video Resume Dataset

In this work, we use a dataset previously collected by our group [116]. Nguyen et al. collected 939 videos using various keywords (like video resume, video cv etc), collected these videos from YouTube. Of these, we randomly selected a subset of 313 videos (i.e. 1/3 of the data) as manual transcriptions is an expensive and time consuming process. Furthermore, of the 313 videos, 21 were discarded due to difficulty in transcription (due to music, accent of speakers) and missing annotations. Hence in this work, we use a corpus of 292 YouTube video resumes. These 292 videos were annotated categorically for gender (Male or Female), ethnicity (Caucasian, Indian, Asian, African or Latin American), and on $1-5$ Likert scale for language proficiency and audio quality of the video resume [116]. 186 videos in the corpus were annotated as "Male". In terms of ethnicity too the corpus is very diverse with 146 annotated as "Caucasian", 69 as "Indians", 42 as "Asian", 13 as "African" and 21 as "Latin American". The mean rated audio quality was 3.6 (min=1.6; max = 5) and an average language proficiency of 3.88 (min=1.6;max=5). This diversity in ethnicity, language proficiency and audio quality indicates the inherent challenge in analysis of this "in-the-wild" corpus.

Annotations

These 292 videos were then further annotated for personality and hirability impressions by AMT workers. Specifically, the AMT workers were asked to rate each video for *Overall Impression, Overall Hirability, Professional Skills, Social Skills, Communication Skills* on a $1-5$ Likert scale and was ensured that each video was rated by at least 5 workers. In this work, we only focus on the hirability variables, the personality impressions will be taken up as future work. Table 7.1 provides details of the variables annotated, their descriptive statistics and their inter rater agreement assessed using Intraclass Correlation Coefficient (ICC). ICC is a commonly used metric in psychology and computing to measure the agreement between raters [147]. Specifically, $ICC(1,k)$ is used as a measure of the inter-rater agreement because the average of 5 randomly selected raters' measurements are used. We observe that the $ICC(1,k)$ values are greater than 0.5 which is considered acceptable in literature.

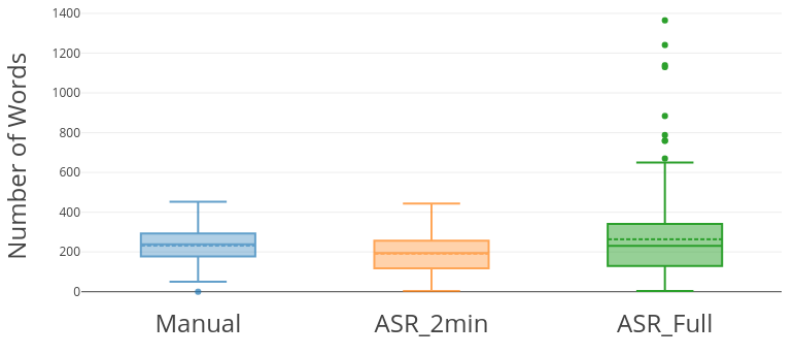


Figure 7.1 – Box plot illustrating the distribution of number of words obtained by (a) manual transcription [Man] (b) ASR for first 2 minutes [ASR-2min] (c) ASR for full video [ASR-Full] for a random subset of 292 videos. The dotted line indicates the mean value.

Transcriptions

To investigate the connections between verbal content and the formation of first impressions, we first transcribed the 292 videos using both manual and automatic methods, as detailed below.

Manual Transcription: To investigate the role of verbal content in the formation of hirability impressions in an ideal case, we manually transcribe the 292 videos. The transcription was carried out by a native English speaker who transcribed the videos as is (with no changes or corrections). As the manual transcription is a tedious and expensive process, only the first *2 minutes* was transcribed. These transcriptions constitute the "gold-standard" as they can be considered as the output of an ideal, error-less automatic speech recognition systems (ASR).

Automatic Speech Recognition: To automatically transcribe the speech-to-text, we used an off-the-shelf ASR Google Speech API [33]. This cloud-based system is based on deep learning techniques for speech transcription [17]. Specifically, this ASR system uses a Long Short-term Memory Recurrent Neural Networks (LSTM RNNs) for speech recognition. This deep neural network based model is shown to outperform the Gaussian Mixture Model (GMM) acoustic models by using "discriminative training", differentiating phonetic units instead of modeling each one independently [139]. Google Speech API was chosen as it ranks among the best performing ASR systems and is readily available [85].

We used this API to generate two sets of transcriptions from the same randomly selected subset of 292 videos (a) first two minutes so as to compare with manual transcription (b) transcription of full videos. We then computed the word error rate (WER), a common measure

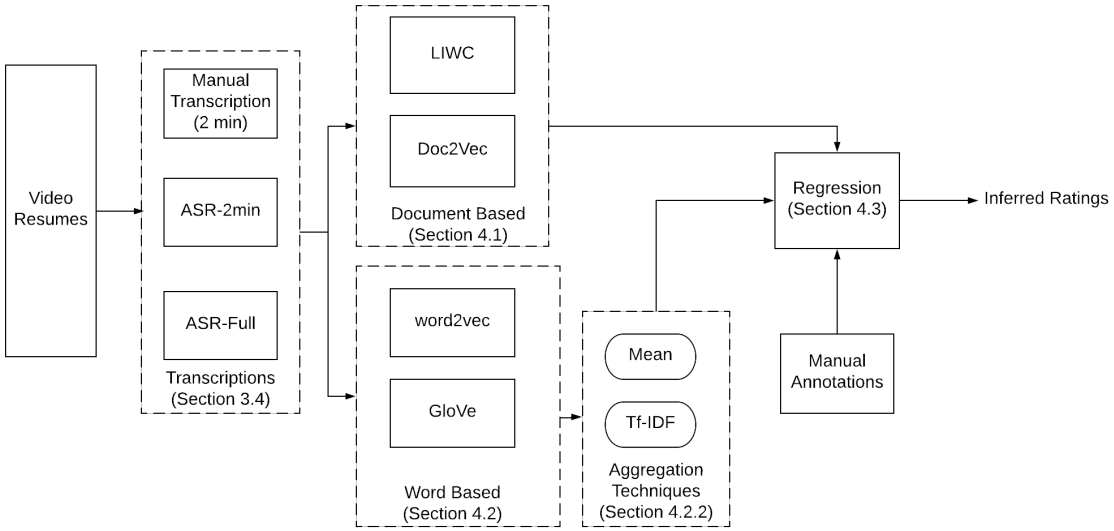


Figure 7.2 – Overview of the work flow used in this study. The two classes of verbal content representation methods (a) document-based (b) word-based investigated is illustrated. For the document-based method, performance of LIWC and Doc2Vec in inferring hirability impressions, individually and combination is investigated. For the word-based method, all combinations of algorithm and aggregation techniques are investigated.

of performance of a speech recognition system. To compute the WER, we compare the reference (manual transcriptions) to the output of the ASR. WER is defined as:

$$WER = \frac{S + D + I}{N} \tag{7.1}$$

where S,D,I and N denote number of substitutions, number of deletions, number of insertions, number of words in the reference respectively.

For this dataset, the WER is 41.5%, indicating that the ASR system performed rather well on this difficult and noisy dataset. One reason for this error rate is that the videos are in conversational style with varied speaking rate and pauses. Due to this, the task of transcribing the speech with accurate punctuation is a difficult one for the ASR system. Another reason for this WER we believe, is the variability in microphone location. In some videos the microphone is close by (close-field) and further away (far-field) which degraded the ASR performance.

In terms of statistics, the manually transcribed data contains a total of 63053 words ($mean$ = 231.8) while the ASR-2min data contains 53917 words ($mean$ = 191.2). For some videos the performance was low with just 4 words being transcribed (99 for manual transcription), while for other cases ASR was almost as good as manual. This difference might be due to reasons inherent to the “in-the-wild” nature of the dataset, including accent of the person, the ambient

noise or low overall audio quality. The total number of words transcribed for the full video is greater (75149) than the manual transcription data. This was expected as the complete video was used for ASR-Full data. Figure 7.1 illustrates the distribution of number of words obtained from the three methods.

To put these results in perspective, Biel et al. reported an WER of 62.4% in their work [20] where the videos were comparable in terms of audio quality. They used the then state-of-the-art system consisting of two two-pass systems that use acoustic models for English based on individual head-mounted microphones (IHM) and single-distant microphone (SDM), respectively [65]. Results obtained using the Google Speech API on standard dataset (TIMIT [57]) showed a mean WER of 9% as compared to Microsoft API (18%) and CMU Sphinx API (37%). This indicates that the video resumes from YouTube was challenging for the ASR, as it is noisy, "in-the-wild" dataset.

7.1.3 Method

The method used in this study is illustrated in Figure 7.2. To obtain a feature representation of verbal content, we evaluated two distinct approaches: (a) representation at the document level; and (b) representation at the word level, followed by an aggregation step.

Document-Based Representation

In this representation method, the entire document (i.e the transcription of one video resume) is used to extract features. This representation has been previously explored in the context of the UBIMPRESSED data corpus. These representations include (a) LIWC and (b) Doc2Vec. The details of these representations has been outlined in Section 3.5 and the same process was followed in this work as well.

Word-Based Representations

The motivation behind studying word based representation was explore the use of words and its context. Specifically, the word-based representation learn the contextually use of certain word forms (like noun, verb). Towards this, we use a two-step approach. First, each word from the transcription is mapped to a word embedding using pre-trained models (word2vec and GloVe). Second, word representations are aggregated to form a document level representation (i.e a feature representation of the transcript).

Word Representation

We used and evaluated the two word representations, Word2Vec and GloVe. Before extracting both of these representations, the corpus was pre-processed by first removing stop words from the text. Then, the text was converted into lower case, stemmed and tokenized using the

Natural Language Toolkit (NLTK) python package [21].

Word2Vec: developed by Mikolov et al., is an unsupervised learning algorithm that learns word embeddings from a text corpus [102]. Two learning models were proposed (a) continuous bag of words (CBOW) (b) continuous skip-gram (skip-gram). In both, the algorithm starts with a randomly initialized vectors and then learns the embeddings by prediction. The CBOW model learns by predicting the current word based on the context (i.e., words around it), while the skip-gram model predicts the surrounding words given the current word. Specifically, the algorithm computes the dot product between the target word and the context words with the an aim to minimize this distance by performing Stochastic Gradient Descent (SGD). When two words are encountered in a similar context, their link, or spacial distance, is reinforced.

In our work, we use the CBOW model of Word2Vec for learning word embeddings. We use the pre-trained vectors provided by Google. These 300-dimensional vectors were obtained by training on the Google News Dataset consisting of 100 billion words and a vocabulary of 3 million words [102].

GloVe: is a statistical method to learn word embeddings developed by Pennington et al. [128]. This algorithm uses the global co-occurrence statistics, i.e count of word co-occurrences in a text corpus. In this work, we use GloVe with two different pre-trained models provided by the authors. (1) GloVe(S) is a 300-dimensional vector trained on 6 billion words of Wikipedia (2014) with a vocabulary size of 400K words. (2) GloVe(B) is a 300-dimensional vector trained on a larger corpus of 840 billion words with a vocabulary of 2.2 million words.

Aggregation Techniques

In order to use Word2Vec and GloVe for representing documents (document embeddings), various aggregation techniques were applied. The most common aggregation techniques are averaging and term frequency - inverse document frequency (TF-IDF). They have been shown in literature to work better than Doc2Vec for short sentences and small documents [84, 38, 166].

Averaging: is the simplest method of aggregation where the document is represented as the average of its word embeddings. This is method takes into account sentence of different length.

Weighted Average: method allows capturing of words that a more valuable than other in a sentence. The simplest method is use of term frequency - inverse document frequency (TF-IDF). TF-IDF is a bag-of-words representation divided by each word's frequency in the document. This method has been shown to perform very well in literature [38, 166].

Regression

We outline the proposed computational framework for evaluating our research questions in a regression task. This task is defined as inferring the impressions of hirability and soft skills using various representations of verbal content. Towards this, we evaluate two regression techniques (Support Vector Machines (SVM) and Random Forest (RF)) implemented in the "scikit-learn" package for Python [126]. The hyperparameters of the machine learning algorithms were optimized for best performance using 10-fold inner cross-validation (CV), while the performance was assessed using the 100 independent runs of Leave-one-video-out CV. The performance of machine learning algorithms was evaluated by two standard measures: coefficient of determination (R^2) and root-mean-square error (RMSE). The baseline performance was $R^2 = 0$, and the inference performance of all the verbal content representation methods was compared against each other to answer of research questions. As RF outperformed SVM in all the inference experiments, we report only the results of RF.

7.1.4 Results and Discussion

In this section, we present the results of the inference experiments. We then discuss these results in the context of our research questions and related work.

RQ-1: Manual Transcriptions

Regression results using manual transcriptions are presented in Table 7.2. We observe that in an ideal case (i.e using manual transcriptions), verbal content explains up to 23% of variance for *Overall Hirability* using the GloVe(S) representation. This observation showcases that rater formed their first impressions of hirability at least partially based on verbal content. These inference results using verbal content are consistently higher than those reported with nonverbal behaviors as predictors [116].

In terms of inference performance, Doc2Vec consistently performs worse for all the hirability variables with $R^2 = 0.08$ (*Overall Hirability*) being highest. One hypothesis to explain these poor results could be the relatively short length of the documents. In this manually transcribed text corpus, the mean number of words is 232.67 (min=50; max=453). As the performance of Doc2Vec is much lower than the other representations, we will not discuss this method thereon.

Overall, competitive results were obtained for LIWC features with highest for *Professional* ($R^2 = 0.24$), followed by *Overall Hirability* ($R^2 = 0.20$), *Overall Impression* and *Communication* (both $R^2 = 0.13$). This indicates that simple features like LIWC captures some of the variance in data. In the context of existing work, these results are better than what is reported in literature. Muralidhar et al., using LIWC features extracted from 169 videos, reported an inference performance of $R^2 = 0.11$ [106]. In another work, Chen et al. reported a correlation of $r = 0.39$ between LIWC only features and expert ratings. For comparison, we convert r to

Table 7.2 – Results of the inference task using the random forest algorithm (N=292) and features extracted from the manually transcribed text corpus as predictors.

	Overall Impress	Overall Hirability	Professional	Social	Communication
LIWC	0.13	0.20	0.24	0.07	0.13
Doc2Vec	0.03	0.08	0.03	0.03	0.05
Word2Vec					
- Avg	0.18	0.17	0.16	0.14	0.22
- TF-IDF	0.20	0.17	0.20	0.10	0.22
GloVe(S)					
- Avg	0.21	0.23	0.17	0.17	0.20
- TF-IDF	0.15	0.23	0.15	0.12	0.15
GloVe(B)					
- Avg	0.12	0.14	0.11	0.14	0.16
- TF-IDF	0.16	0.19	0.13	0.13	0.15

R^2 by squaring the values, and obtain $R^2 = 0.15$. In both these works, *face-to-face interviews* in a lab-setting was used for data collected and manually transcriptions to create the text corpus. Using 1891 video interviews, Chen et al [30] obtained *Precision* and *Recall* of 0.67 and 0.66 respectively in a classification task. The authors obtained the text corpus using ASR provided by IBM Bluemix platform and representation was achieved using Bag-of-Words (BoW). In this work, the authors used *video interviews* as data source, with Amazon Mechanical Turk workers playing the role of participants. Nguyen et al., collected 939 conversational video resumes from YouTube and investigate the impact of nonverbal behavior in inferring first impressions. They reported a inference performance of $R^2 = 0.20$ for *Social* and *Communication*, $R^2 = 0.15$ for *Overall Hirability*.

Both word-based representation, Word2Vec and GloVe yielded competitive results, independent of the aggregation methods. Comparatively, GloVe(B) had lower inference performance than GloVe(S) and Word2Vec, with best performance of $R^2 = 0.19$ for *Overall Hirability* with TF-IDF aggregation. For the same aggregation method, the lowest performance was for *Social* and *Professional* with $R^2 = 0.13$. Using the averaging aggregation the performance was a little lower, with $R^2 = 0.14$ (highest) for *Overall Hirability*, $R^2 = 0.11$ (lowest) for *Professional*.

The GloVe(S) algorithm using averaging technique performed best amongst all the representation for all hirability variables except *Professional*. The best performance was achieved for *Overall Hirability* with $R^2 = 0.23$ followed by *Overall Impression* ($R^2 = 0.21$), *Communication* ($R^2 = 0.20$) with lowest for Professional and Social ($R^2 = 0.17$). It is interesting to note that GloVe(S) model performed better than GloVe(B) which was pre-trained on a much larger data.

The Word2Vec representation performed better than LIWC features for *Overall Impression, Social* and *Communication*, but slightly lower for *Overall Hirability* and *Professional*. For this representation, the TF-IDF aggregation technique performed better then simple averaging

Table 7.3 – Results of the inference task using the random forest algorithm. (N=292). The manually transcribed (Manual) and automatically transcribed (ASR-2min) text corpus were used as predictors.

	Overall Impress		Overall Impress		Professional		Social		Communication	
	Manual	ASR-2min	Manual	ASR-2min	Manual	ASR-2min	Manual	ASR-2min	Manual	ASR-2min
LIWC	0.13	0.13	0.20	0.17	0.24	0.18	0.07	0.11	0.13	0.17
Word2Vec - TF-IDF	0.20	0.18	0.17	0.21	0.20	0.26	0.10	0.13	0.22	0.19
GloVe(S) - Avg	0.21	0.14	0.23	0.12	0.17	0.12	0.17	0.13	0.20	0.07
GloVe(B) - TF-IDF	0.16	0.12	0.19	0.11	0.13	0.07	0.13	0.10	0.15	0.09

(Avg) for *Overall Impression* ($R^2 = 0.2$ and $R^2 = 0.18$) and *Professional* ($R^2 = 0.20$ and $R^2 = 0.16$). The inference performance for the two aggregation techniques was same for *Overall Hirability* ($R^2 = 0.17$) and *Communication* ($R^2 = 0.22$).

Overall, in terms of aggregation techniques, we cannot observe a consistent improvement from averaging to TF-IDF. Rather, we observe that averaging (Avg) yields slightly higher performance when used in conjunction with GloVe(S) ($R^2 = 0.23$ for *Overall Hirability*). This tendency is reversed with GloVe(B) and Word2Vec. In summary, for manually transcribed text corpus, GloVe(S) model (with averaging) achieved the best inference performance with for all hirability variables except *Professional* (for which LIWC was the best) followed by Word2Vec model with TF-IDF aggregation method. Our results indicate the improved performance of word-based representations of verbal content in inferring hirability impressions, thus answering **RQ-1**.

RQ-2: Effect of Automatic Transcriptions

Regression results obtained for ASR-2min text corpus and compare them to those obtained for manually transcribed text (Manual) in Table 7.3. The ASR-2min text corpus is obtained by using the first 2 min of the video resumes (same duration transcribed manually) automatically transcribed using the Google Speech API. We observe that in this text corpus, verbal content explains up to 21% of variance for *Overall Hirability* using the Word2Vec representation. These inference results using verbal content are slightly higher than those reported with nonverbal behaviors are predictors [116].

From the table we observe that LIWC features extracted from the manual transcriptions perform slightly better than those extracted from using the ASR-2min text corpus in the inference tasks except for *Social* ($R^2 = 0.07$ compared to $R^2 = 0.11$) and *Communication* ($R^2 = 0.13$ compared to $R^2 = 0.17$). Specifically, we observe that features extracted from Manual perform slightly better than those from ASR-2min for *Overall Hirability* ($R^2 = 0.20$ compared to $R^2 = 0.17$) and *Professional* ($R^2 = 0.24$ compare to $R^2 = 0.18$). Interesting ASR-2min LIWC features perform slightly better for *Social* ($R^2 = 0.11$) and *Communicative* ($R^2 = 0.17$).

An interesting observation is that GloVe(S) model that had the highest inference performance

for Manual does not perform as well with the ASR-2min text corpus. It performs worse for *Communication* ($R^2 = 0.07$) (compared to $R^2 = 0.20$) and best for *Overall Impression* with $R^2 = 0.14$ (compared to $R^2 = 0.21$ for Manual). Similarly, GloVe(B) model performs worse than other models individually and in comparison with results from Manual with $R^2 = 0.12$ for *Overall Impression*, $R^2 = 0.11$ for *Overall Hirability*. This model performs worse for *Professional* ($R^2 = 0.07$) followed by *Communication* ($R^2 = 0.09$).

The best performing model using ASR-2min text corpus is Word2Vec with TF-IDF aggregation techniques. This model performs best for *Professional* ($R^2 = 0.26$) followed by *Overall Hirability* ($R^2 = 0.21$), *Communication* ($R^=0.19$), *Overall Impression* ($R^2 = 0.18$) and performs worse for *Social* ($R^2 = 0.13$). We then compare this performance with that of the same model using manual transcriptions. We observe that except for *Communication* and *Overall Impression*, use of ASR-2min performs slightly better than manual transcriptions. We hypothesis that this improvement in performance for some hirability variables is due to Word2Vec model having a greater vocabulary size (3M words) and the weighted averaging (TF-IDF) aggregation techniques.

A similar comparative study was conducted by Biel et al. [20] who investigated the use of manual and automatic transcription to infer personality impressions in in YouTube video blogs (vlogs). The authors reported a much lower performance using ASR ($R^2 = 0.18$) as compared to manual transcriptions $R^2 = 0.31$ for *Agreeableness*. This can be attributed to the high WER (62.4%) of the ASR system the authors used [65].

In summary, our results indicate that inference performance of ASR-2min is comparable to Manual albeit with a different representation, indicating the feasibility of using deep neural network based ASR for transcriptions in conjunction with advances in natural language processing (like Word2Vec and GloVe) for verbal content analysis, in understanding the connections between words used and formation of first impressions, thus answering **RQ-2**.

RQ-3: Effect of Duration

In this subsection, we present the inference results obtained using the ASR-Full text corpus and compare them with those obtained for the ASR-2min text corpus (Table 7.4). The ASR-Full corpus contains the entire duration of the 292 video resumes, automatically transcribed using the Google Speech API. The total number of words in the ASR-Full corpus is 75149 ($mean = 276.8$) while the same numbers for ASR-2min is 53917 words ($mean = 191.2$). We hypothesis that this increase in average number of words will aid in improved inference performance.

Best Inference performance using LIWC features extracted from ASR-Full text corpus was obtained for *Communication* and *Professional* with $R^2 = 0.20$, followed by *Overall Hirability* ($R^2 = 0.19$). The worst performance was for *Social* with $R^2 = 0.09$. Comparing these values with inference results obtained ASR-2min indicates that transcription of the extra duration seem to

Table 7.4 – Results of the inference task using the random forest algorithm. Both the automatically transcribed text corpus, ASR-2min and ASR-Full were used as predictors. (N=292)

	Overall Impression		Overall Hirability		Professional		Social		Communication	
	ASR-2min	ASR-Full	ASR-2min	ASR-Full	ASR-2min	ASR-Full	ASR-2min	ASR-Full	ASR-2min	ASR-Full
LIWC	0.13	0.14	0.17	0.19	0.18	0.20	0.11	0.09	0.17	0.20
Word2Vec										
- Avg	0.09	0.16	0.08	0.22	0.17	0.13	0.09	0.22	0.14	0.21
- TF-IDF	0.18	0.16	0.21	0.16	0.26	0.10	0.13	0.19	0.19	0.14
Glove(S)										
- Avg	0.14	0.19	0.12	0.14	0.12	0.09	0.13	0.11	0.07	0.12
- TF-IDF	0.14	0.20	0.14	0.18	0.11	0.14	0.15	0.10	0.09	0.16
Glove(B)										
- Avg	0.13	0.16	0.09	0.12	0.06	0.12	0.11	0.14	0.08	0.16
- TF-IDF	0.12	0.11	0.11	0.08	0.07	0.09	0.10	0.12	0.09	0.13

improve inference performance. Specifically, there is slight improvement in performance for all variables expect Social ($R^2 = 0.09$ compared to $R^2 = 0.11$) with the largest improvement for *Communication* ($R^2 = 0.20$ compared to $R^2 = 0.17$), followed by *Overall Hirability* ($R^2 = 0.19$ compared to $R^2 = 0.17$). We believe this improvement is solely due to increased mean number of words in each transcription.

For word-based representations, we observe that both
Word2Vec(Avg) and GloVe(TF-IDF) yielded competitive results.
Word2Vec(Avg) performed better than Word2Vec(TF-IDF) method for all social variables with best performances for *Overal Hirability* and *Social* with $R^2 = 0.22$ and worse for *Professional* ($R^2 = 0.13$). Using this representation method, ASR-Full out-performed the ASR-2min corpus for all variables except *Professional* ($R^2 = 0.17$ compared to $R^2 = 0.13$).

Inference performance of GloVe(S) with TF-IDF aggregation
performed slightly better than the averaging method for all variables, with best performance achieved for *Overall Impression* with $R^2 = 0.20$, followed by *Overall Hirability* ($R^2 = 0.18$) and worse for *Social* ($R^2 = 0.10$). This method also performed better using the full dataset (ASR-Full) when compared to ASR-2min with improved inference for all variables except *Social* ($R^2 = 0.10$ as compare to $R^2 = 0.15$).

The GloVe(B) performed worse of all the representations with $R^2 = 0.16$ (*Overall Impression, Communication*) being the best and worst for *Overall Hirability* and *Professional* ($R^2 = 0.12$). Although the performance of GloVe(B) method was worse that other word-based representations, it showed improved inference performance when compared with features obtained from ASR-2min.

In summary, we observe a moderate improvement in inference performance with increase in duration of video transcribed using LIWC and GloVe(S). This validates our hypothesis and answers **RQ3**. Again it must be noted that these inference performance is comparable to those obtained using manual transcriptions ("gold standard") and is better than those reported using nonverbal cues as predictors [116].

7.1.5 Conclusion

This work investigated the relationship between verbal content and the formation of first impressions in conversational video resumes. To this end, we use 292 noisy "in-the-wild" video resumes from YouTube, previously collected by Nguyen et al. [116]. We investigated the effect of manual versus automatic transcriptions on inference performance, using various document-based and word-based representations as features.

Regarding **RQ1**, we observed that LIWC, GloVe(S) and Word2Vec representation yielded competitive inference performance. The best performance achieved using GloVe(S) method (with averaging) ($R^2 = 0.23$ for *Overall Hirability*), which higher than those reported by Nguyen et al. [116] using nonverbal behaviors as predictors ($R^2 = 0.15$ for *Overall Hirability*).

Regarding **RQ2**, we observed that inference performance of ASR as almost as good as those obtained using manual transcriptions with Word2Vec representation as predictors. Specifically, using ASR, the best inference performance of $R^2 = 0.26$ (*Professional*) followed by $R^2 = 0.21$ for *Overall Hirability* was obtained as compared to $R^2 = 0.20$ and $R^2 = 0.18$ respectively using manual transcriptions.

Regarding **RQ3**, we observed an improved inference performance for all the representation methods using the entire duration of the video for ASR. The best performance was obtained using Word2Vec (Avg) method with $R^2 = 0.22$ for *Overall Impression* and *Social*. This representation method also outperformed the ASR-2min corpus for all hirability variables except *Professional* with $R^2 = 0.17$ (ASR-2min) compared to $R^2 = 0.13$ (ASR-Full).

There are certain limitations of this work. First, the data corpus studied is small which has lead to inference performance using Doc2Vec despite using pre-trained models. Second, there is no one best way to represent verbal content with the best representation for manually transcribed corpus being GloVe(S) with averaging aggregation, while those for automatic transcriptions are Word2Vec(TF-IDF) for ASR-2min and GloVe(S) with TF-IDF aggregation for ASR-Full. Another limitation of this work is that the use of deep neural network based representations, though improve inference performance, does not aid in identifying specific verbal cues and hence cannot be used in behavioral training systems. The low performance of verbal features could be due to the inherent difference in language used the participant (i.e people choose different words to express themselves based on their education levels and socioeconomic status [69]). Furthermore, word representation is an ongoing research topic in NLP. Improved representation techniques and larger data corpus might aid in better inference performance using verbal features.

In summary, the results of **RQ1** indicate the feasibility of using advances in natural language processing (like Word2Vec and GloVe) for verbal content representation. The results of **RQ2** and **RQ3** indicate the feasibility of using deep neural network based ASR for transcriptions in understanding the connections between words used and formation of first impressions in online video resumes from YouTube. In future, we will analyze and compare the nonverbal

behavioral cues for the this data corpus (292 videos). We will also access the impact of verbal content on first impression using the complete dataset of 939 video resumes.

7.2 Dites-Moi: Wearable Feedback on Conversational Behavior

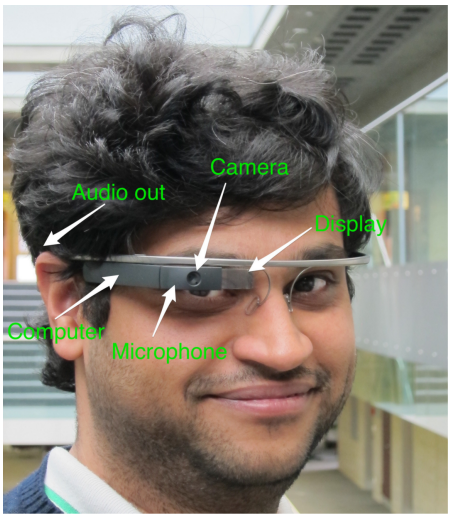
Existing psychology literature indicates that social interaction skills can be improved by practicing both verbal and nonverbal communication including how much, how fast, and how loud to talk, and how to regulate turn taking [71]. Advances in ubiquitous and wearable computing are enabling new possibilities to deliver real-time feedback [64, 119] and uses in the classroom, like physics experiments [165].

Providing real-time feedback during conversations has been investigated in the past. In the context of group interactions, feedback to participants was provided by projecting their speaking time on a large common surface like a wall [41] or on a customized table which acted as both sensing and display platform [10]. A mobile phone-based solution for sensing and displaying a person's nonverbal cues (speaking time, prosody and body movements) was developed in [86], yielding a reduction of behavioral differences between dominant and non-dominant participants. In [156], feedback systems that combine visual and acoustic cues, e.g. automatically estimating speaking time and visual attention using headbands tracked by infrared camera were developed.

In the context of public speaking, Google Glass has been used as a head-mounted display system to provide real-time feedback on a presenter's posture openness, body energy, and speech rate sensed using data provided by Kinect and an external microphone [37]. In [158], Google Glass has been used to display information and as an audio sensor to provide automatic real-time feedback on a speaker's speaking rate and energy. The data was processed on an external server.

In the context of dyadic interaction, [97] investigated the effect of a head mounted device on social interaction. The authors reported a degradation of social interaction and eye contact. However, display on the screen was a series of slides showing emails, text messages etc with each slide being visible for 40 seconds. We see an opportunity in the design of tools using Google Glass that provide real-time feedback during face-to-face interaction to increase self-awareness of conversational behavior, while not impairing the quality of interaction. To our knowledge, little work has been done on the implementation of a training procedure using Google Glass that provides real-time feedback during face-to-face conversation to increase self-awareness of basic nonverbal cues, while not impairing the quality of interaction.

The objective of this work was to evaluate an automatic, real-time, conversational behavior awareness system for young sales apprentices to make them aware of their nonverbal behavior while interacting with a customer. Towards this objective, we utilized a pilot Google Glass app, designed and developed by Jean Costa, a PhD student at Cornell University who was a collaborator in this project. This app provides real-time behavioral feedback and its design



(a) Glass Sensors

(b) Speak More

(c) Speak Less

Figure 7.3 – Overview of Google Glass sensors and visual feedback on Google Glass Display

and usefulness was evaluated by conducting a pilot study with 15 sales apprentices from a VET school. We believe that some of the results obtained in this study can also be applied to other devices capable of displaying behavioral feedback, such as smart watches, tablets, etc.

The contributions of this work are as follows. First, we evaluated a Google Glass app for providing real-time feedback on speaking time. Second, we collected a dataset of 15 sales interactions with apprentices recorded in the lab, where Google Glass was used as a self-awareness tool. Third, we demonstrated the usefulness of this tool and analyzed the effect of real-time feedback on conversation. Our work constitutes a first step towards designing a real-time behavioral training tool in a dyadic setting that is appealing to young apprentices in jobs involving positive interpersonal skills.

The material in this section was originally published in [105].

7.2.1 Approach

Providing feedback during a dyadic interaction without negative impact is a challenge. The main challenge is to provide speakers with key behavioral insights without distracting them from their interaction. The human brain is not adept at multitasking [123], hence any significant distraction might lead to behavioral artifacts like stuttering, awkward pauses or smiles. Additionally, by continuously staring at the feedback screen, the speaker might lose eye contact with the protagonist, causing the quality of interaction to degrade.

Considering these constraints, Google Glass was utilized as a display to provide feedback. The Google Glass screen is a small high resolution display located at the periphery of user's

Figure 7.4 – View of the study setting: the participant works behind the desk; the interaction partner is not visible in the figure.

field of vision and engineered to have minimal cognitive load. For aural feedback, the bone conduction transducer was utilized. Using this technology, audio is sent directly to the inner ear through bones of the skull, rendering it audible only to the user.

The behavioral tool consists of two main components; sensing and feedback. The sensing component is responsible for perceiving and processing the user's nonverbal behavior. The built-in microphone of Google Glass was exploited towards this end (Fig. 1a). The feedback generation component uses the resulting analysis and presents the appropriate messages either visually or aurally. The following section details both the components of our behavioral awareness tool and an user case study conducted to verify the effectiveness of the said tool.

App Implementation

A prototype of the behavioral awareness tool was designed and implemented on two Google Glass devices using the Android platform. The devices were running Android OS and implementation of the application was done in Java.

A significant effort was devoted to evaluate what nonverbal features should be shared with the user. Literature in psychology and social computing indicate several nonverbal cues to be important during a dyadic interactions in the context of workplaces [87, 114]. However, due to constraints of low computational power of Google Glass, nonverbal cues with moderate computational requirements were considered. In an initial design phase, speaking time and a proxy of head orientation were considered and implemented. Hence, feasibility of estimated gaze and speaking time was investigated.

A small pre-trial was conducted with three lab colleagues to evaluate the experience of this initial design. The results of the pilot study showed that the use of the two nonverbal cues (speaking time and gaze) significantly affected the duration of Google Glass battery and lead to heating of the device to uncomfortable levels. Another reason that necessitated a simple interface was the sample population in the evaluation use case. Participants in the user study reported themselves to be inexperienced with wearable devices (mean= 1.8, median= 2) and reluctant to use new technology (mean= 3.1, median= 3) (scale $1-7$). Given these factors, the final design was focused on speaking time, which is intuitive to users engaged in conversation, and is backed up by literature in psychology as a cue related to extraversion and dominance among other constructs [87].

To compute speaking status, speech captured by the built-in microphone of Google Glass was utilized. The speech non-speech segmentation was performed using a two-step approach. First, the subject's voice was segmented from the other protagonist using audio energy as a discriminative feature: the microphone is significantly closer to the subject than the other interlocutor, therefore the subject's voice is assumed to be louder. Second, we used the method proposed by [14], which was shown to be robust in noisy environments [93]. The method is independent to energy and uses a two-layer binary HMM: the low-level latent variable is voiced/non-voiced and the high-level one is speech/non-speech. The processing was done on Google Glass itself.

The sensing component provides analysis for the feedback component. Our prototype currently furnishes feedback based on a window of 20 seconds. If no voice activity is detected for 20 seconds, Google Glass prompts the user to speak. If continuous voice activity is detected for more than 20 seconds, the tool prompts the user to stop talking. Although we acknowledge that this 20-second threshold can be somewhat arbitrary, this duration was chosen based on detailed discussions with colleagues in psychology and its use in existing literature [158]. Additionally, the objective of this study was not to investigate the best speaking duration; rather, we focused our analysis on the effect of feedback on the quality of the interaction.

Feedback on speaking time was provided as one of two possible modalities: visual and aural. For both modalities, the feedback was not noticeable by the other interlocutor. Visual feedback was provided using text, which was presented to the user sparsely as suggested by authors in [158]. The text, screenshot in Figures 7.3b & 7.3c, prompted the participants to '*speak less*' or to '*speak more*'.

Aural feedback, a modality that has been used less often in the social sensing literature, was provided in the form of prerecorded speech ('*speak less*','*speak more*'). The bone conduction transducer was utilized to provide this feedback.

Table 7.5 – List of questions in the self reported pre- and post-questionnaires rated on a Likert scale

Pre-Questions (Self Reported)	Post-Questions (Self Reported)	
	Visual	Audio
Sales Experience	Usefulness	Usefulness
Google Glass Experience	Distracting	Distracting
Interest in Google Glass	Overall Impression	Overall Impression
Interest in Technology		

Scenario

To evaluate the usefulness of the system, we conducted a user study with 15 participants. Subjects were volunteers from a local VET school, who participated as an opportunity to improve their communication and sales skills. Of the 15 students, 9 were male. Average age was 17.7 years old. The subjects reported little professional sales experience (mean= 1.75, median= 1 on a $1-7$ Likert scale). They were randomly split with one half provided with visual feedback while the other group was presented with audio feedback.

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 salesperson role. Each student had to interact with a customer with the goal to satisfy the client, and try to sell them the best (= most expensive) data package along with the phone (iPhone). During the interaction, Google Glass would provide automatic feedback on behavioral cues. It was in their discretion to follow the suggestion or not. The role of the client was played by a researcher who was a native French speaker with directions to elicit two behaviors from the participants: talk for a relatively long time, and remain silent. A snapshot of the scenario is presented in Figure 7.4. All interactions (for both partners) and Google Glass feedback are video recorded with Kinect devices (Fig. 2).

7.2.2 Evaluation

To gain insight on number of times feedback was given and the type of feedback provided, the videos were manually coded by the authors. Due to the design of the experiment, each participant received feedback at least once. It was further observed that three participants received feedback twice. Also, three participants received feedback to '*speak more*, while other received feedback '*speak less*'.

To understand the user's perspective on using Google Glass, the app, and to identify issues with current prototype implementation, evaluation was carried out by analysis of participant self-reported questionnaires and external annotator impressions.

Figure 7.5 – Distribution of participants' self ratings for (a) feedback modalities (higher is better for *useful* and *Ov. Impression*, lower is better for *Distract*)

Questionnaire Data

The participants were asked to fill two questionnaires, one before and the other after the interaction. In both questionnaires, subjects had to rate various questions on a Likert scale (where 1 ='very poor'; 7 ='very good'). The list of questions asked in both the questionnaires are presented in Table 7.5. Additionally, the pre-questionnaire consisted of demographic details and a personality test. The personality test, in French, was administered using a Ten Item Personality Inventory (TIPI) [62]. Analysis of personality is planned as part of future work. The post-questionnaire required subjects to answer only for the modality they were presented during the data collection.

Figure 7.5 shows the self rated impressions of participants for both modalities of feedback. It can be observed that participants find the feedback using audio modality to be useful (median = 4) but find it to be a little distracting (median = 3). On the other hand, participants who were given visual feedback found this modality to be less distracting than audio (median = 2) and more useful (median = 5). A possible explanation for this difference is that the audio feedback could have interfered with the speech of the protagonist. The participants reported a positive overall impression for both modalities (median = 5).

Broadly, the participants indicated a positive overall impression towards an wearable behavioral feedback tool (Figure 7.6). They also indicated that the wearable device and app were found to be Natural (median = 3.5), Cool (median = 5.5), Comfortable (median = 5) and Fun (median = 5) during the dyadic interaction. Thus, the results indicate that subjects found the real-time behavioral feedback to be useful, natural and comfortable. These results are in line with those reported in literature [37, 158], and are novel from the perspective of the specific use of Google Glass by a very young population. At the end of data collection, in an informal discussion with the participants, they all expressed the usefulness of the app. In particular, participants favored visual feedback over audio feedback.

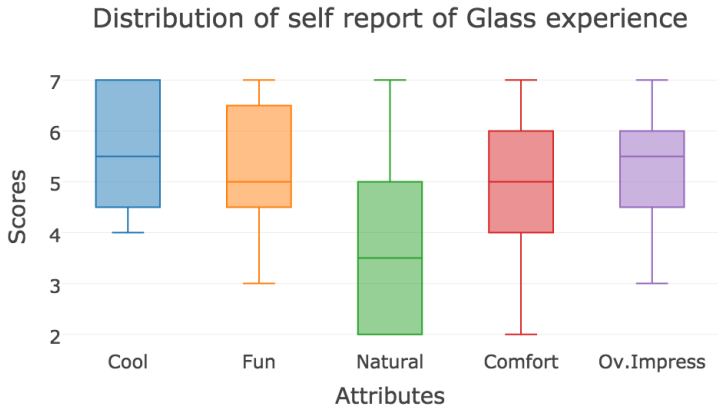


Figure 7.6 – Distribution of participants' self ratings for overall experience (higher is better)

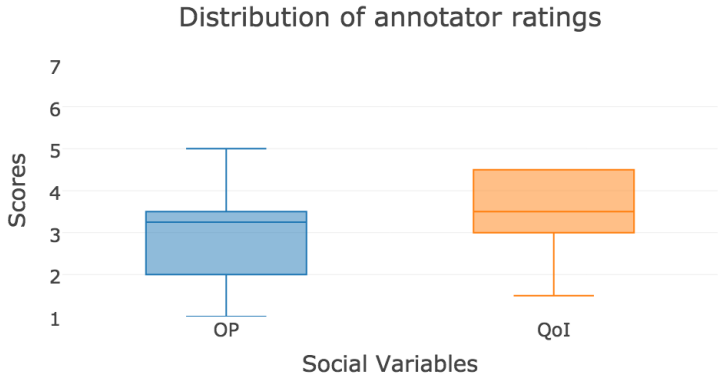


Figure 7.7 – Distribution of participants' self ratings for *overall performance* and *quality of interaction* ratings by annotators (higher is better)

External Observer Annotations

To assess the impact of glass on dyadic interaction, the sales video was 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 Google Glass 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* (a) *Overall performance (OP) of the participant* (b) *Quality of interaction (QoI) with the participant.*

Agreement between the raters in Group-B was calculated using Intraclass Correlation Coefficient (*ICC*), as a measure commonly used in psychology and social computing [147]. $ICC(2,k)$ was used as all raters gave scores for each video. The obtained ICC values were above 0.70 for

Table 7.6 – 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.

Feedback Type	Reaction	Time to heed
Speak Less	Smiling, Laughing, Squinting	1-4 seconds
Speak More	Smiling	2-4 seconds

both the social variables [OP: $ICC(2,k) = 0.90$, $p < .001$; QoI: $ICC(2,k) = 0.70$, $p < .001$]. This indicates that the agreement between raters was high for both social variables. Final scores for both social variables were obtained by taking the mean of all scores.

The distribution of annotation data for both variables is presented in Figure 7.7. 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 7.8 indicates that for majority of the videos, the annotators of Group-A were unable to correctly infer if feedback had been provided (combining the “no” and “maybe” columns in Figure 7.8). These results suggest that in several cases the reaction of Google Glass 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 7.6. This in conjunction with annotations by Group-A on inference of feedback (Figure 7.8) indicate that in the majority of the cases reaction to feedback was natural.

7.2.3 Conclusion

This work presented the design and evaluation of a real-time wearable prototype for self-awareness of conversational behavior, aim to support young VET students. Towards this end, we assessed an Android-based app on Google Glass. Speaking time was chosen to give feedback based on existing literature. The tool was evaluated in a study consisting of a newly collected corpus of 15 students from local VET school in a dyadic sales pitch scenario. The evaluation of questionnaire data provided insights about usefulness and distraction of the tool during a social interaction. An interesting observation has been the positive acceptance of glass by this age group in contrast to poor acceptance of Google Glass in general. This could

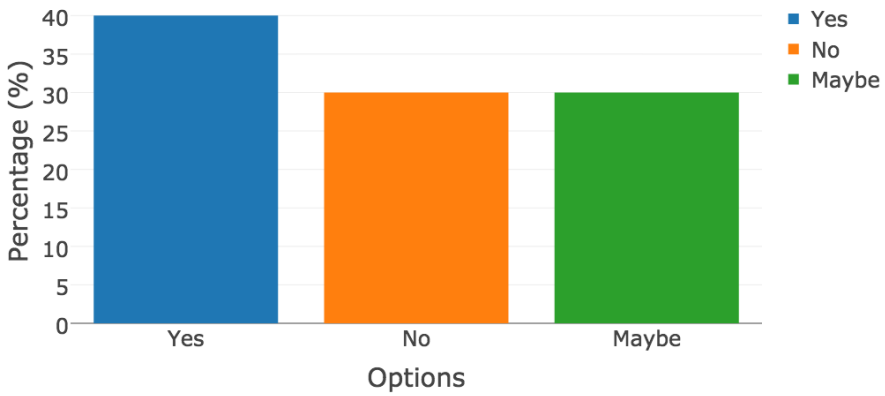


Figure 7.8 – Distribution of answers for prediction by Group-A. All Google Glass users received feedback.

be due to novelty of the device or the fact that this generation may be more accepting of new technologies such as Google Glass and similar devices as they are exposed to technology from an younger age. We believe this would be an interesting area to be explored in future. We also plan to explore the use of multiple nonverbal features for feedback including eye gaze or number of pauses. The challenges are to sense and process data by offloading intensive computation to a phone, without hindering dyadic interaction.

8 Conclusions

In this thesis, we developed a computational framework, based on the principles of unobtrusive measurements and social signal processing, for automatic recognition of first impressions in two dyadic interactions in hospitality settings. Towards this, we contributed to the collection of a new dataset (known as UBIMPRESSED) consisting of 169 simulated dyadic interactions in two hospitality settings, job interviews and reception desk (total of 338 interactions). The UBIMPRESSED data corpus is the result of a behavioral training framework designed and developed for students in an international hospitality school to improve their first impression in workplace settings. To understand the formation of first impressions and to quantify behavior, we extracted various verbal and nonverbal features automatically. We then studied each setting separately.

In Chapter 4, we compiled the results of our investigation of the job interview setting using verbal and nonverbal behavioral cues. A Pearson's correlation analysis between the automatically extracted cues and various perceived social variables revealed interesting gender differences, which are backed by psychology literature. A computational framework for automatic inference of first impressions achieved $R^2 = 0.32$ using nonverbal cues, while verbal cues obtained from manual transcription of videos, revealed low predictive performance ($R^2 = 0.11$). Furthermore, we observed that the fusion of verbal and nonverbal cues improved inference performance ($R^2 = 0.34$). We conducted a principal component analysis to investigate a lower dimensional representation of the soft skills annotated. The first principal component was associated with overall positive and negative impressions. This components, when used as labels in a regression task, achieved a performance of $R^2 = 0.41$.

In Chapter 5, we presented the results of our investigation between perceived job performance, Big-5 personality trait impressions, perceived attractiveness and automatically extracted verbal and nonverbal cues at the reception desk setting. A Pearson's correlation analysis showed a positive connection between participants who spoke for a longer duration, faster, took longer turns, and had fewer silence events obtained higher scores for performance and skill impressions. Interestingly, we also found connections between the nonverbal behavioral displayed by the client. Specifically, clients tend to speak faster with greater visual motion

in presence of receptionists who were rated higher. We also found a negative correlation between attractiveness attributes like *likeable* and *friendly* and *performance impression* scores for women participants, while there was no correlation for men. An automatic computational framework to infer perceived performance using audio-visual nonverbal cues achieved R^2 = 0.30, while the Big-5 trait impressions achieve a performance of R^2 = 0.35 The best inference performance was obtained by a fusing of NVB cues and Big-5 impressions (R^2 = 0.37).

In Chapter 6, we investigated the effect of settings on the formation of first impressions and human behavior across two situations; job interview and reception desk. Specifically, our objective was to infer perceived job performance on the job using verbal and nonverbal behavior displayed during job interviews. In the first step, we conducted Pearson's correlation analysis between cross-situation variables. We observed moderate correlations between perceived variables in job interviews and perceived performance and soft skills on the job. The correlation analysis also revealed that some behavioral cues (greater speaking turn duration and head nods) were positively correlated to higher ratings for all perceived variables independent of the situation. We then evaluated a computational framework for inferring perceived performance and soft skills on the job. This framework utilized perceived variables in job interviews automatically extracted verbal and nonverbal cues from both job interviews and reception desk. Our best inference model, a fusion of nonverbal cues extracted from the reception desk interaction and the human-rated interview scores, achieved R^2 = 0.40. While using verbal cues represented by a fusion of LIWC and Doc2Vec features achieved an inference performance of R^2 = 0.25 for perceived performance.

In Chapter 7, we presented two additional application of our work. First, we discussed our work, investigating the choice of words during self-presentation and its connections to first impressions. Towards this, we used 292 noisy, "in-the-wild" video resumes from YouTube, which is a subset of a previously collected dataset by Nguyen et al. [**?**]. We investigated the (a) various representations of verbal content by leveraged existing NLP methods, (b) effect of using automatic speech recognition on the inference performance. In the first step, we had the video resumes transcribed both manually and automatically (using cloud-based Google Speech API). Next, we extracted various representations of verbal content (LIWC, Doc2Vec, Word2Vec, and GloVe) to be used as features. We then evaluated the inference performance in a regression task for various representation using manually transcribed data and obtained the best performance of R^2 = 0.23 using the GloVe model for *Overall Hirability*. Use of automatically transcribed data in an inference task achieved similar performance for *Overall Hirability* (R^2 = 0.21) .

The second application was the development and evaluation of a real-time feedback system using advances in wearable computing. The system provided behavior awareness during dyadic interaction using Google Glass. We evaluated the effectiveness of this real-time feedback system in a pilot study consisting of 15 apprentices from a local vocational training school. The Google Glass provided real-time information to the participant about his/her speaking duration, a behavioral cue which has been shown to impact the formation of first

impressions. Overall, participants found the real-time feedback system fun, little distracting and useful. Furthermore, analysis of the manual coding of the interaction videos showed no negative influence of the wearable sensing and real-time feedback on the dyadic social interaction

8.1 Implications

We believe that the insights from our work have implications for hospitality and other customer-facing domains where interpersonal communication and soft skills are critical. Here, we discuss implications of our work in the domains of hospitality, psychology, and ubiquitous computing.

In the hospitality industry, there is great emphasis on soft skills and interpersonal communication as they are considered critical to business [157, 54]. Our work contributes to this domain by showing connections between the automatically extracted nonverbal behavior of potential employees (displayed during a job interview), and their perceived performance on the job. As the objective of job interviews is to select the best candidate for a given position, our work provides important inferences for human resource teams and hiring managers in the domain of hospitality. Our research also shows the feasibility of utilizing an automatic framework using both verbal and nonverbal behavior for assessing candidates for customer-facing roles. Given that we have identified some of the most important nonverbal behaviors in the job interview that are moderately connected to future job performance (speaking longer, louder, with fewer silences, gesturing more while speaking and nodding for longer periods of time), training in hospitality and other service-related fields might put an emphasis on students learning these behaviors and maintaining them also under stressful conditions (e.g., dissatisfied clients).

In psychology, our work is a step in the direction of understanding human behavior in multiple situations by integrating ubiquitous computing and social psychology. Our research shows the importance of face-to-face job interviews for their predictive value for perceived performance in customer-facing jobs. Recruiters underscore the importance of a personal meeting with a job applicant and often talk about the importance of "feeling" the applicant. This "feeling" might refer to observing the applicant's nonverbal behavior and one's own nonverbal reactions to it. This is in line with our findings indicating that some of the applicant's and the interviewer's nonverbal cues have predictive power on the perceived performance on the job (reception desk) [40, 104].

Finally, in ubiquitous computing, our work has implications for developing behavioral training systems. Specifically, the observation that the same behavioral cues are positively linked to first impressions across settings encourages the development of behavioral awareness systems that focus on specific cues. Such systems could be helpful for individuals who aspire to improve the nonverbal behavior they convey [72, 53]. Such systems are also potentially important to socially challenged individuals to express and/or perceive nonverbal communication. Overall, understanding differences in behavior across situations and the information they convey

is important to build ubiquitous computational devices capable of sensing and responding unobtrusively [121, 164].

8.2 Limitations & Future Work

There are some limitations to our work.

1. The size of our data corpus. The UBIMPRESSED dataset consists of 169 simulated job interviews and reception desk interactions making a total of 338 interactions. To our knowledge, this constitutes one of the largest academic datasets of work-related dyadic interactions. Even this number is relatively small to explore other inference methods including recent advances in deep neural networks (DNNs). Furthermore, this corpus consists of 94 females and 75 males, which is relatively small to explore the weak connection observed during our gender analysis.

2. Our computational models are trained on human annotation of impressions. These impression scores carry the biases of the raters and hence the computational models learned from them. This thesis has used 5 raters for interviews and 3 for the desk all of whom where Caucasians. We believe that the use of a more diverse sample of raters, both in terms of gender and ethnicity would help reduce the biases learned by the computational models.

3. The results of this thesis, which indicate the feasibility of utilizing an automatic framework for assessing candidates for customer-facing roles does not imply that the hiring process can be fully automated. We believe that automation should be a tool with human supervision. This is need to ensure any bias learned by the model does not lead to any discrimination dues to gender or ethnicity.

4. There are limitations of the behavioral features examined in this thesis. As one of the objectives of this work is to provide feedback, we focused on extracting behavioral features that are interpretable. There are many low-level audio and visual features which have not been examined as they are not human interpretable. For example, use of MFCC features in speech is universal, they are not easily understood or can be used to modify one's behavior. Use of such features may improve inference performance. Similarly, use of deep neural network (DNN) many help improve inference performance, but the DNN based representations may not be interpretable. The results of this thesis, especially the use of DNN based NLP representation methods, will enable the exploration of other behavioral cues (both DNN-based and other low-level features) in both audio and video modality.

5. Although this thesis has been investigating hospitality workplace situations, we believe that some of the reported results can be generalized to other settings. Specifically, the results of the cross-situation analysis in Chapter 6 can be generalized to service industry situations which need high degree of interpersonal communication skills. We believe this model will not be applicable to other jobs that need high technical skills and low interpersonal skills like programmers or systems engineers, and will be a separate topic of research in future.

6. The use of verbal content in video resumes has some limitations. This work too suffers from a small data size due to which we observed that Doc2Vec representation was not effective. Another limitation is that the video we investigated were of the same language, use of multi-lingual text data has been explored and shown to be feasible in NLP. Another limitation in this work is the moderately high word error rate due to the conversational style of the videos.

7. The Google Glass based feedback system has been evaluated using a small data corpus and single nonverbal behavioral cue. Other cues and devices will need to be explored to provide unobtrusive feedback system. An important research question would be the acceptability of constant feedback from Google Glass. In our personal experience, constant reminders by devices in other settings like Google maps providing constant instructions while driving leads to increase irritability and stress. Hence, the acceptability of constant reminders of ones behavior will be an important area to be explored in future.

Bibliography

[1] http://www.dev-audio.com/products/microcone/.

[2] AHEARNE, M., GRUEN, T. W., AND JARVIS, C. B. If looks could sell: Moderation and mediation of the attractiveness effect on salesperson performance. *Int. J. Research in Marketing 16*, 4 (1999), 269–284.

[3] AMALFITANO, J. G., AND KALT, N. C. Effects of eye contact on the evaluation of job applicants. *Journal of Employment Counseling 14*, 1 (1977), 46–48.

[4] AMBADY, N., HALLAHAN, M., AND ROSENTHAL, R. On judging and being judged accurately in zero-acquaintance situations. *J. Personality and Social Psychology 69*, 3 (1995).

[5] AMBADY, N., HALLAHAN, M., AND ROSENTHAL, R. On judging and being judged accurately in zero-acquaintance situations. *J. Personality and Social Psychology 69*, 3 (1995).

[6] AMBADY, N., KRABBENHOFT, M. A., AND HOGAN, D. The 30-sec sale: Using thin-slice judgments to evaluate sales effectiveness. *J. Consumer Psychology 16*, 1 (2006), 4–13.

[7] AMBADY, N., AND ROSENTHAL, R. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin 111*, 2 (1992).

[8] AMBADY, N., AND ROSENTHAL, R. Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *J. Personality and Social Psychology 64*, 3 (1993), 431.

[9] AMBADY, N., AND SKOWRONSKI, J. J. *First impressions.* Guilford Press, 2008.

[10] BACHOUR, K., KAPLAN, F., AND DILLENBOURG, P. Reflect: An interactive table for regulating face-to-face collaborative learning. In *European Conference on Technology Enhanced Learning* (2008), Springer, pp. 39–48.

[11] BARNUM, C., AND WOLNIANSKY, N. Taking cues from body language. *Management Review 78*, 6 (1989), 59–61.

[12] BARRICK, M. R., AND MOUNT, M. K. The big five personality dimensions and job performance: a meta-analysis. *Personnel Psychology 44*, 1 (1991), 1–26.

[13] BARRICK, M. R., SHAFFER, J. A., AND DEGRASSI, S. W. What you see may not be what you get: relationships among self-presentation tactics and ratings of interview and job performance. *J. Applied Psychology 94*, 6 (2009), 1394.

[14] BASU, S. A linked-hmm model for robust voicing and speech detection. In *Proc. of ICASSP* (2003), vol. 1, IEEE.

[15] BATRINCA, L. M., MANA, N., LEPRI, B., PIANESI, F., AND SEBE, N. Please, tell me about yourself: automatic personality assessment using short self-presentations. In *Proc. ACM ICMI* (2011).

[16] BAYES, M. A. Behavioral cues of interpersonal warmth. *J. Consulting and Clinical Psychology 39*, 2 (1972), 333.

[17] BEAUFAYS, F. The neural networks behind google voice transcription. *Google Research blog* (2015).

[18] BIEL, J.-I., AND GATICA-PEREZ, D. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Trans. on Multimedia 15*, 1 (2013).

[19] BIEL, J.-I., AND GATICA-PEREZ, D. Mining crowdsourced first impressions in online social video. *IEEE Transactions on Multimedia 16*, 7 (2014), 2062–2074.

[20] BIEL, J.-I., TSIMINAKI, V., DINES, J., AND GATICA-PEREZ, D. Hi youtube!: Personality impressions and verbal content in social video. In *Proc. 15th ACM ICMI* (2013), ACM, pp. 119–126.

[21] BIRD, S., KLEIN, E., AND LOPER, E. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc, 2009.

[22] BONACCIO, S., O'REILLY, J., O'SULLIVAN, S. L., AND CHIOCCHIO, F. Nonverbal behavior and communication in the workplace: A review and an agenda for research. *J. Management 42*, 5 (2016), 1044–1074.

[23] BONO, J. E., AND ILIES, R. Charisma, positive emotions and mood contagion. *The Leadership Quarterly 17*, 4 (2006), 317–334.

[24] BORKENAU, P., AND LIEBLER, A. Trait inferences: Sources of validity at zero acquaintance. *J. Personality and Social Psychology 62*, 4 (1992), 645.

[25] BREIMAN, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.

[26] BURGOON, J. K., BIRK, T., AND PFAU, M. Nonverbal behaviors, persuasion, and credibility. *Human Communication Research 17*, 1 (1990), 140–169.

[27] CANEEL, R. *Social signaling in decision making.* PhD thesis, Massachusetts Institute of Technology, 2005.

[28] CHEN, L., FENG, G., LEONG, C. W., LEHMAN, B., MARTIN-RAUGH, M., KELL, H., LEE, C. M., AND YOON, S.-Y. Automated scoring of interview videos using doc2vec multimodal feature extraction paradigm. In *Proc. 18th ACM ICMI* (2016), ACM, pp. 161–168.

[29] CHEN, L., YOON, S.-Y., LEONG, C. W., MARTIN, M., AND MA, M. An initial analysis of structured video interviews by using multimodal emotion detection. In *Proc. Workshop Emotion Recognition in the Wild Challenge and Workshop* (2014), ACM, pp. 1–6.

[30] CHEN, L., ZHAO, R., LEONG, C. W., LEHMAN, B., FENG, G., AND HOQUE, M. E. Automated video interview judgment on a large-sized corpus collected online. In *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on* (2017), IEEE, pp. 504–509.

[31] CHEN, Y., YU, Y., AND ODOBEZ, J.-M. Head nod detection from a full 3d model. In *Proc. IEEE ICCV Workshops* (2015).

[32] CHUNG, C., AND PENNEBAKER, J. W. The psychological functions of function words. *Social Communication* (2007), 343–359.

[33] CLOUD SERVICES, G. Google Speech API.

[34] COGNITIVE SERVICES, M. Azure Emotion API.

[35] CORTES, C., AND VAPNIK, V. Support-vector networks. *Machine learning 20*, 3 (1995), 273–297.

[36] CURHAN, J. R., AND PENTLAND, A. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *J. Applied Psychology 92*, 3 (2007).

[37] DAMIAN, I., TAN, C. S. S., BAUR, T., SCHÖNING, J., LUYTEN, K., AND ANDRÉ, E. Augmenting social interactions: Realtime behavioural feedback using social signal processing techniques. In *Proc. ACM CHI* (2015).

[38] DE BOOM, C., VAN CANNEYT, S., DEMEESTER, T., AND DHOEDT, B. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters 80* (2016), 150–156.

[39] DEGROOT, T., AND GOOTY, J. Can nonverbal cues be used to make meaningful personality attributions in employment interviews? *J. Business and Psychology 24*, 2 (2009).

[40] DEGROOT, T., AND MOTOWIDLO, S. J. Why visual and vocal interview cues can affect interviewers' judgments and predict job performance. *J. Applied Psychology 84*, 6 (1999), 986.

[41] DiMicco, J. M., Pandolfo, A., and Bender, W. Influencing group participation with a shared display. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work* (2004), ACM, pp. 614–623.

[42] Dion, K. K. Cultural perspectives on facial attractiveness.

[43] Dovidio, J. F., and Ellyson, S. L. Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening. *Social Psychology Quarterly* (1982), 106–113.

[44] Eagly, A. H., Ashmore, R. D., Makhijani, M. G., and Longo, L. C. What is beautiful is good, but...: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological bulletin 110*, 1 (1991), 109.

[45] End, C. M., and Saunders, K. Short communication: Powerless and jobless? comparing the effects of powerless speech and speech disorders on an applicant's employability. *Frontiers 2*, 1 (2013).

[46] Erickson, B., Lind, E. A., Johnson, B. C., and O'Barr, W. M. Speech style and impression formation in a court setting: The effects of "powerful" and "powerless" speech. *J. Experimental Social Psychology 14*, 3 (1978), 266–279.

[47] Forbes, R. J., and Jackson, P. R. Non-verbal behaviour and the outcome of selection interviews. *J. Occupational Psychology 53*, 1 (1980).

[48] Forbes, R. J., and Jackson, P. R. nonverbal behaviour and the outcome of selection interviews. *J. Occupational Psychology 53*, 1 (1980), 65–72.

[49] Frauendorfer, D., and Mast, M. S. The impact of nonverbal behavior in the job interview. *The Social Psychology of Nonverbal Communication* (2014), 220.

[50] Frauendorfer, D., Mast, M. S., Nguyen, L., and Gatica-Perez, D. Nonverbal social sensing in action: Unobtrusive recording and extracting of nonverbal behavior in social interactions illustrated with a research example. *J. Nonverbal Behavior 38*, 2 (2014).

[51] Funder, D. C. Towards a resolution of the personality triad: Persons, situations, and behaviors. *J. Research in Personality 40*, 1 (2006), 21–34.

[52] Funes-Mora, K. A., and Odobez, J.-M. Gaze estimation in the 3d space using rgb-d sensors. *International Journal of Computer Vision 118*, 2 (2016), 194–216.

[53] Fung, M., Jin, Y., Zhao, R., and Hoque, M. E. Roc speak: semi-automated personalized feedback on nonverbal behavior from recorded videos. In *Proc. ACM UBIComp* (2015), ACM, pp. 1167–1178.

[54] Gabbott, M., and Hogg, G. An empirical investigation of the impact of non-verbal communication on service evaluation. *European J. Marketing 34*, 3/4 (2000), 384–398.

[55] GABBOTT, M., AND HOGG, G. The role of non-verbal communication in service encounters: A conceptual framework. *J. Marketing Management 17*, 1-2 (2001), 5–26.

[56] GARIMELLA, V. R. K., ALFAYAD, A., AND WEBER, I. Social media image analysis for public health. In *Proc. 2016 CHI Conf. on Human Factors in Computing Systems* (2016), ACM, pp. 5543–5547.

[57] GAROFOLO, J. S., LAMEL, L. F., FISHER, W. M., FISCUS, J. G., AND PALLETT, D. S. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n 93* (1993).

[58] GATICA-PEREZ, D. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing 27*, 12 (2009).

[59] GATICA-PEREZ, D. Signal processing in the workplace. *IEEE Signal Process. Mag. 32*, 1 (2015).

[60] GIFFORD, R., NG, C. F., AND WILKINSON, M. Nonverbal cues in the employment interview: Links between applicant qualities and interviewer judgments. *J. Applied Psychology 70*, 4 (1985), 729.

[61] GORN, G. J., GOLDBERG, M. E., AND BASU, K. Mood, awareness, and product evaluation. *J. Consumer Psychology 2*, 3 (1993), 237–256.

[62] GOSLING, S. D., RENTFROW, P. J., AND SWANN, W. B. A very brief measure of the big-five personality domains. *J. Research in Personality 37*, 6 (2003), 504–528.

[63] GUÉGUEN, N., AND JACOB, C. Clothing color and tipping: Gentlemen patrons give more tips to waitresses with red clothes. *J. Hospitality & Tourism Research 38*, 2 (2014), 275–280.

[64] HA, K., CHEN, Z., HU, W., RICHTER, W., PILLAI, P., AND SATYANARAYANAN, M. Towards wearable cognitive assistance. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services* (2014), ACM, pp. 68–81.

[65] HAIN, T., BURGET, L., DINES, J., GARNER, P. N., GRÉZL, F., EL HANNANI, A., HUIJBREGTS, M., KARAFIAT, M., LINCOLN, M., AND WAN, V. Transcribing meetings with the amida systems. *IEEE Transactions on Audio, Speech, and Language Processing 20*, 2 (2012), 486–498.

[66] HAMERMESH, D. S., AND PARKER, A. Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review 24*, 4 (2005), 369–376.

[67] HENNIG-THURAU, T., GROTH, M., PAUL, M., AND GREMLER, D. D. Are all smiles created equal? how emotional contagion and emotional labor affect service relationships. *J. Marketing 70*, 3 (2006), 58–73.

[68] HIEMSTRA, A. *Fairness in paper and video resume screening.* 2013.

[69] HOFF, E., AND TIAN, C. Socioeconomic status and cultural influences on language. *Journal of communication Disorders 38*, 4 (2005), 271–278.

[70] HOLLANDSWORTH, J. G., GLAZESKI, R. C., AND DRESSEL, M. E. Use of social-skills training in the treatment of extreme anxiety and deficient verbal skills in the job-interview setting. *J. Applied Behaviour Analysis 11*, 2 (1978).

[71] HOLLANDSWORTH, J. G., KAZELSKIS, R., STEVENS, J., AND DRESSEL, M. E. Relative contributions of verbal, articulative, and nonverbal communication to employment decisions in the job interview setting. *J. Personnel Psychology 32*, 2 (1979).

[72] HOQUE, M. E., COURGEON, M., MARTIN, J.-C., MUTLU, B., AND PICARD, R. W. Mach: My automated conversation coach. In *Proc. ACM UBICOMP* (2013).

[73] HOWARD, J. L., AND FERRIS, G. R. The employment interview context: Social and situational influences on interviewer decisions. *J. Applied Social Psychology 26*, 2 (1996), 112–136.

[74] HUFFCUTT, A. I., CONWAY, J. M., ROTH, P. L., AND STONE, N. J. Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *J. Applied Psychology 86*, 5 (2001).

[75] HUNG, H., AND GATICA-PEREZ, D. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia 12*, 6 (2010), 563–575.

[76] HUNG, H., HUANG, Y., FRIEDLAND, G., AND GATICA-PEREZ, D. Estimating dominance in multi-party meetings using speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing 19*, 4 (2011), 847–860.

[77] IMADA, A. S., AND HAKEL, M. D. Influence of nonverbal communication and rater proximity on impressions and decisions in simulated employment interviews. *J. Applied Psychology 62*, 3 (1977).

[78] JACOB, C., GUÉGUEN, N., BOULBRY, G., AND ARDICCIONI, R. Waitresse's facial cosmetics and tipping: A field experiment. *Int. J. Hospitality Management 29*, 1 (2010), 188–190.

[79] JAYAGOPI, D. B., HUNG, H., YEO, C., AND GATICA-PEREZ, D. Modeling dominance in group conversations using nonverbal activity cues. *Audio, Speech, and Language Processing, IEEE Trans. on 17*, 3 (2009), 501–513.

[80] JOHNSON, S. K., PODRATZ, K. E., DIPBOYE, R. L., AND GIBBONS, E. Physical attractiveness biases in ratings of employment suitability: Tracking down the "beauty is beastly" effect. *The J. Social Psychology 150*, 3 (2010), 301–318.

[81] JUNG, H. S., AND YOON, H. H. The effects of nonverbal communication of employees in the family restaurant upon customer's emotional responses and customer satisfaction. *Int. J. Hospitality Management 30*, 3 (2011), 542–550.

[82] KANG, J., AND HYUN, S. S. Effective communication styles for the customer-oriented service employee: Inducing dedicational behaviors in luxury restaurant patrons. *Int. J. Hospitality Management 31*, 3 (2012), 772–785.

[83] KENRICK, D. T., AND FUNDER, D. C. Profiting from controversy: Lessons from the person-situation debate. *American psychologist 43*, 1 (1988), 23.

[84] KENTER, T., BORISOV, A., AND DE RIJKE, M. Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640* (2016).

[85] KËPUSKA, V., AND BOHOUTA, G. Comparing speech recognition systems (microsoft api, google api and cmu sphinx). *Int. J. Eng. Res. Appl 7* (2017), 20–24.

[86] KIM, T., CHANG, A., HOLLAND, L., AND PENTLAND, A. S. Meeting mediator: enhancing group collaboration with sociometric feedback. In *CHI'08 extended abstracts on Human factors in computing systems* (2008), ACM, pp. 3183–3188.

[87] KNAPP, M., HALL, J., AND HORGAN, T. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[88] KUHN, M. A short introduction to the caret package.

[89] LE, Q., AND MIKOLOV, T. Distributed representations of sentences and documents. In *Proc. 31st ICML* (2014), pp. 1188–1196.

[90] LEAPER, C., AND ROBNETT, R. D. Women Are More Likely Than Men to Use Tentative Language, Aren't They? A Meta-Analysis Testing for Gender Differences and Moderators. *Psychology of Women Quarterly 35*, 1 (2011).

[91] LEIGH, T. W., AND SUMMERS, J. O. An initial evaluation of industrial buyers' impressions of salespersons' nonverbal cues. *Journal of Personal Selling & Sales Management 22*, 1 (2002), 41–53.

[92] LEPRI, B., MANA, N., CAPPELLETTI, A., AND PIANESI, F. Automatic prediction of individual performance from thin slices of social behavior. In *Proc. 17th ACM MMProc.* (2009), ACM, pp. 733–736.

[93] LU, H., FRAUENDORFER, D., RABBI, M., MAST, M. S., CHITTARANJAN, G. T., CAMPBELL, A. T., GATICA-PEREZ, D., AND CHOUDHURY, T. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proc. ACM UBICOMP* (2012).

[94] LUOH, H.-F., AND TSAUR, S.-H. Physical attractiveness stereotypes and service quality in customer–server encounters. *The Service Industries Journal 29*, 8 (2009), 1093–1104.

[95] MADAN, A., CANEEL, R., AND PENTLAND, A. S. Groupmedia: distributed multi-modal interfaces. In *Proc. 6th International Conference on Multimodal Interfaces* (2004), ACM, pp. 309–316.

[96] MANO, H., AND OLIVER, R. L. Assessing the dimensionality and structure of the consumption experience: evaluation, feeling, and satisfaction. *J. Consumer research 20*, 3 (1993), 451–466.

[97] MCATAMNEY, G., AND PARKER, C. An examination of the effects of a wearable display on informal face-to-face communication. In *Proc. of the SIGCHI Conf. on Human Factors in computing systems* (2006), pp. 45–54.

[98] MCGOVERN, T. V. *The making of a job interviewee: The effect of nonverbal behavior on an interviewer's evaluations during a selection interview.* PhD thesis, ProQuest Information & Learning, 1977.

[99] MCGOVERN, T. V., AND TINSLEY, H. E. Interviewer evaluations of interviewee nonverbal behavior. *J. Vocational Behavior 13*, 2 (1978).

[100] MEHL, M. R., ROBBINS, M. L., AND GROSSE DETERS, F. Naturalistic observation of health-relevant social processes: the electronically activated recorder (ear) methodology in psychosomatics. *Psychosomatic Medicine 74*, 4 (2012), 410.

[101] MICZO, N., SEGRIN, C., AND ALLSPACH, L. E. Relationship between nonverbal sensitivity, encoding, and relational satisfaction. *Communication Reports 14*, 1 (2001), 39–48.

[102] MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S., AND DEAN, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (2013), pp. 3111–3119.

[103] MOK, C., SPARKS, B., AND KADAMPULLY, J. *Service quality management in hospitality, tourism, and leisure.* Routledge, 2013.

[104] MOTOWIDLO, S. J., AND BURNETT, J. R. Aural and visual sources of validity in structured employment interviews. *Organizational Behavior and Human Decision Processes 61*, 3 (1995), 239–249.

[105] MURALIDHAR, S., COSTA, J. M. R., NGUYEN, L. S., AND GATICA-PEREZ, D. Dites-Moi : Wearable Feedback on Conversational Behavior. In *Proc. 15th ACM MUM* (2016).

[106] MURALIDHAR, S., AND GATICA-PEREZ, D. Examining Linguistic Content and Skill Impression Structure for Job Interview Analytics in Hospitality. In *Proc. 16th ACM MUM* (2017).

[107] MURALIDHAR, S., MAST, M. S., AND GATICA-PEREZ, D. A tale of two interactions: Inferring performance in hospitality encounters from cross-situation social sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2*, 3 (2018), 129.

[108] MURALIDHAR, S., NGUYEN, L. S., FRAUENDORFER, D., ODOBEZ, J.-M., SCHIMD-MAST, M., AND GATICA-PEREZ, D. Training on the Job: Behavioral Analysis of Job Interviews in Hospitality. In *Proc. 18th ACM ICMI* (2016), pp. 84–91.

[109] MURALIDHAR, S., NGUYEN, L. S., AND GATICA-PEREZ, D. Words Worth: Verbal Content and Hirability Impressions in YouTube Video Resumes. In *Proc. of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (2018).

[110] MURALIDHAR, S., SCHIMD-MAST, M., AND GATICA-PEREZ, D. How May I Help You? Behavior and Impressions in Hospitality Service Encounters. In *Proc. 19th ACM ICMI* (2017).

[111] MURALIDHAR, S., SIEGFRIED, R., ODOBEZ, J.-M., AND GATICA-PEREZ, D. Facing Employers and Customers: What Do Gaze and Expressions Tell About Soft Skills? In *Proc. 17th ACM MUM* (2018).

[112] NAIM, I., TANVEER, M. I., GILDEA, D., AND HOQUE, M. E. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. *Proc. IEEE FG* (2015).

[113] NGUYEN, L. S. *Computational Analysis Of Behavior In Employment Interviews And Video Resumes*. PhD thesis, École Polytechnique Fédérale de Lausanne, May 2015.

[114] NGUYEN, L. S., FRAUENDORFER, D., MAST, M. S., AND GATICA-PEREZ, D. Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Trans. on Multimedia 16*, 4 (2014).

[115] NGUYEN, L. S., AND GATICA-PEREZ, D. I would hire you in a minute: Thin slices of nonverbal behavior in job interviews. In *Proc. ACM ICMI* (2015).

[116] NGUYEN, L. S., AND GATICA-PEREZ, D. Hirability in the wild: Analysis of online conversational video resumes. *IEEE Transactions on Multimedia 18*, 7 (2016), 1422–1437.

[117] NGUYEN, L. S., MARCOS-RAMIRO, A., MARRÓN ROMERA, M., AND GATICA-PEREZ, D. Multimodal analysis of body communication cues in employment interviews. In *Proc. ACM ICMI* (2013).

[118] NIST, N. Machine translation evaluation official results, 2012.

[119] OFEK, E., IQBAL, S. T., AND STRAUSS, K. Reducing disruption from subtle information delivery during a conversation: mode and bandwidth investigation. In *Proc. ACM CHI* (2013).

[120] OLIVOLA, C. Y., EUBANKS, D. L., AND LOVELACE, J. B. The many (distinctive) faces of leadership: Inferring leadership domain from facial appearance. *The Leadership Quarterly 25*, 5 (2014), 817–834.

[121] PANTIC, M., PENTLAND, A., NIJHOLT, A., AND HUANG, T. S. Human computing and machine understanding of human behavior: A survey. In *Artifical Intelligence for Human Computing*. Springer, 2007, pp. 47–71.

[122] PARSONS, C. K., AND LIDEN, R. C. Interviewer perceptions of applicant qualifications: A multivariate field study of demographic characteristics and nonverbal cues. *J. Applied Psychology 69*, 4 (1984), 557.

[123] PASHLER, H. Dual-task interference in simple tasks: data and theory. *Psychological bulletin 116*, 2 (1994).

[124] PEARCE, W. B., AND CONKLIN, F. Nonverbal vocalic communication and perceptions of a speaker.

[125] PEARSON, K. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and J. Science 2*, 11 (1901), 559–572.

[126] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., AND DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[127] PENNEBAKER, J. W., AND KING, L. A. Linguistic styles: language use as an individual difference. *J. Personality and Social Psychology 77*, 6 (1999), 1296.

[128] PENNINGTON, J., SOCHER, R., AND MANNING, C. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014), pp. 1532–1543.

[129] PENTLAND, A. Social dynamics: Signals and behavior. In *Int. Conf. on Developmental Learning* (2004), vol. 5.

[130] PENTLAND, A., AND HEIBECK, T. *Honest signals: how they shape our world*. MIT press, 2010.

[131] PENTLAND, A. S. Social signal processing [exploratory dsp]. *Signal Processing Magazine, IEEE 24*, 4 (2007), 108–111.

[132] PIANESI, F., MANA, N., CAPPELLETTI, A., LEPRI, B., AND ZANCANARO, M. Multimodal recognition of personality traits in social interactions. In *Proc. ACM ICMI* (2008).

[133] RADUCANU, B., VITRIA, J., AND GATICA-PEREZ, D. You are fired! nonverbal role analysis in competitive meetings. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE Int. Conf. on* (2009), IEEE, pp. 1949–1952.

[134] RASMUSSEN, K. G. Nonverbal behavior, verbal behavior, resumé credentials, and selection interview outcomes. *J. Applied Psychology 69*, 4 (1984), 551.

[135] ŘEHŮŘEK, R., AND SOJKA, P. Software Framework for Topic Modelling with Large Corpora. In *Proc. LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA.

[136] RINIOLO, T. C., JOHNSON, K. C., SHERMAN, T. R., AND MISSO, J. A. Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *The J. General Psychology 133*, 1 (2006), 19–35.

[137] ROTHMANN, S., AND COETZER, E. P. The Big Five Personality Dimensions and Job Performance. *J. Industrial Psychology 29*, 1 (2003).

[138] RUST, R. T., AND VERHOEF, P. C. L. oliver,(1994),"service quality insights and managerial implications from the frontier", 81.

[139] SAK, H., SENIOR, A., AND BEAUFAYS, F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth annual conference of the international speech communication association* (2014).

[140] SANCHEZ-CORTES, D., ARAN, O., MAST, M. S., AND GATICA-PEREZ, D. Identifying emergent leadership in small groups using nonverbal communicative cues. In *Int Conf. on Multimodal interfaces and the workshop on machine learning for Multimodal Interaction* (2010), ACM, p. 39.

[141] SANCHEZ-CORTES, D., ARAN, O., MAST, M. S., AND GATICA-PEREZ, D. A nonverbal behavior approach to identify emergent leaders in small groups. *Multimedia, IEEE Trans. on 14*, 3 (2012), 816–832.

[142] SANCHEZ-CORTES, D., BIEL, J.-I., KUMANO, S., YAMATO, J., OTSUKA, K., AND GATICA-PEREZ, D. Inferring mood in ubiquitous conversational video. In *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia* (2013), ACM, p. 22.

[143] SANCHEZ-CORTES, D., MOTLICEK, P., AND GATICA-PEREZ, D. Assessing the impact of language style on emergent leadership perception from ubiquitous audio. In *Proc. 11th Int. Conf. on MUM.* (2012), ACM, p. 33.

[144] SCHERER, K. R. Methods of research on vocal communication: Paradigms and parameters. *Handbook of Methods in Nonverbal Behavior Research* (1982), 136–198.

[145] SCHLENKER, B. R. *Impression management: The self-concept, social identity, and interpersonal relations*. Brooks/Cole Publishing Company Monterey, CA, 1980.

[146] SCHMIDT, A. Cloud-based ai for pervasive applications. *IEEE Pervasive Computing 15*, 1 (2016), 14–18.

[147] SHROUT, P. E., AND FLEISS, J. L. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin 86*, 2 (1979).

[148] SIEGFRIED, R., YU, Y., AND ODOBEZ, J.-M. Towards the use of social interaction conventions as prior for gaze model adaptation. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (2017), ACM, pp. 154–162.

[149] SIEGMAN, A. W. The telltale voice: Nonverbal messages of verbal communication.

[150] SINGH, V. K., GHOSH, S., AND JOSE, C. Toward multimodal cyberbullying detection. In *Proc. 2017 CHI Conf. Extended Abstracts on Human Factors in Computing Systems* (2017), ACM, pp. 2090–2099.

[151] SÖDERLUND, M., AND JULANDER, C.-R. Physical attractiveness of the service worker in the moment of truth and its effects on customer satisfaction. *J. Retailing and Consumer Services 16*, 3 (2009), 216–226.

[152] SÖDERLUND, M., AND ROSENGREN, S. Dismantling" positive affect" and its effects on customer satisfaction: an empirical examination of customer joy in a service encounter. *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior 17* (2004), 27.

[153] SOLE, D. A. *Getting Started with the Computer Vision API.* Apress, Berkeley, CA, 2018.

[154] SOLE, D. A. *Introducing Microsoft Cognitive Services.* Apress, Berkeley, CA, 2018.

[155] SPISAK, B. R., GRABO, A. E., ARVEY, R. D., AND VAN VUGT, M. The age of exploration and exploitation: Younger-looking leaders endorsed for change and older-looking leaders endorsed for stability. *The Leadership Quarterly 25*, 5 (2014), 805–816.

[156] STURM, J., HERWIJNEN, O. H.-V., EYCK, A., AND TERKEN, J. Influencing social dynamics in meetings through a peripheral display. In *Proceedings of the 9th international conference on Multimodal interfaces* (2007), ACM, pp. 263–270.

[157] SUNDARAM, D., AND WEBSTER, C. The role of nonverbal communication in service encounters. *J. Services Marketing 14*, 5 (2000), 378–391.

[158] TANVEER, M. I., LIN, E., AND HOQUE, M. E. Rhema: A real-time in-situ intelligent interface to help people with public speaking. In *Proc. ACM IUI* (2015), ACM.

[159] TAUSCZIK, Y. R., AND PENNEBAKER, J. W. The psychological meaning of words: Liwc and computerized text analysis methods. *J. Language and Social Psychology 29*, 1 (2010), 24–54.

[160] THORNDIKE, E. L. A constant error in psychological ratings. *Journal of applied psychology 4*, 1 (1920), 25–29.

[161] VAN RIJSBERGEN, C. J. A new theoretical framework for information retrieval. In *ACM SIGIR Forum* (1986), vol. 21, ACM.

[162] VINCIARELLI, A., SALAMIN, H., AND PANTIC, M. Social signal processing: Understanding social interactions through nonverbal behavior analysis. In *CVPR Workshops* (2009), IEEE, pp. 42–49.

[163] VISWESVARAN, C., AND ONES, D. S. Perspectives on models of job performance. *Int. J. Selection and Assessment 8*, 4 (2000), 216–226.

[164] WEISER, M. The computer for the 21st century. *IEEE Pervasive Computing 1*, 1 (2002), 19–25.

[165] WEPPNER, J., HIRTH, M., KUHN, J., AND LUKOWICZ, P. Physics education with google glass gphysics experiment app. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication* (2014), ACM, pp. 279–282.

[166] YIH, W.-T., TOUTANOVA, K., PLATT, J. C., AND MEEK, C. Learning discriminative projections for text similarity measures. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning* (2011), Association for Computational Linguistics, pp. 247–256.

Skanda **Muralidhar**

Rue des Finettes 36, Martigny CH-1920

(+41) 792352585 | skanda.muralidhar@idiap.ch | skanda.muralidhar

Education

Ph.D in Electrical Engineering — *Lausanne, Switzerland*

EPFL(École Polytechnique Fédérale de Lausanne) — *Dec. 2014 - PRESENT*

- Thesis Topic: *Inferring First Impressions in Hospitality Encounters from Cross-Situation Social Sensing*
- Keywords: *Human Computer Interaction, Multimodal Sensing, Ubiquitous Computing, Wearable Sensors*
- Adviser: Professor Daniel GATICA-PEREZ

Masters in Computing — *Singapore*

School of Computing, National University of Singapore — *Aug. 2012 - Oct 2014*

- Thesis Topic: *Human affective state estimation*
- Adviser: Professor Mohan KANKANHALLI

Bachelors in Electrical Engineering — *Bangalore, India*

University Visvesvaraya College Of Engineering (UVCE) — *Sept. 2001 - June 2005*

- *Magna cum Laude*, First Class With Distinction
- Thesis Topic: *Digital Forensics: An Image Processing Approach*

Publications & Patents

Mobile and Ubiquitous Multimedia (MUM) - 2018 — *Cairo, Egypt*

Muralidhar, Skanda and Siegfried, Remy and Odobez, Jean-Marc and Gatica-Perez, Daniel — *Nov 2018*

- Facing Employers and Customers: What Do Gaze and Expressions Tell About Soft Skills?. In: *Proc. of the 17th Int Conf. on Mobile and Ubiquitous Multimedia*, 2018.

Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis at EMNLP - 2018 — *Brussels, Belgium*

Muralidhar, Skanda and Nguyen, Laurent Son and Gatica-Perez, Daniel. — *Oct - Nov 2018*

- Words Worth: Verbal Content and Hirability Impressions in YouTube Video Resumes. In: *In Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* 2018.

PACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT) - 2018 — *Singapore*

Muralidhar, Skanda and Mast, Marianne Schmid and Gatica-Perez, Daniel. — *Sept. 2018*

- A Tale of Two Interactions: Inferring Performance in Hospitality Encounters from Cross-Situation Social Sensing. In:*Proc of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2018.

International Conference on Multimodal Interaction (ICMI) - 2017 — *Glasgow, Scotland*

Muralidhar, Skanda and Mast, Marianne Schmid and Gatica-Perez, Daniel — *Nov 2017*

- How May I Help You? Behavior and Impressions in Hospitality Service Encounters. In: *Proc. of the 19th ACM on Int Conf. on Multimodal Interaction*, 2017.

Mobile and Ubiquitous Multimedia (MUM) - 2017 — *Stuttgart, Germany*

Muralidhar, Skanda and Gatica-Perez, Daniel — *Nov 2017*

- Examining Linguistic Content and Skill Impression Structure for Job Interview Analytics in Hospitality. In: *Proc. of the 16th Int Conf. on Mobile and Ubiquitous Multimedia*, 2017.

International Conference on Multimodal Interaction (ICMI) - 2016 — *Tokyo, Japan*

Muralidhar, Skanda and Nguyen, Laurent Son and Frauendorfer, Denise and Odobez, Jean-Marc and Mast, Marianne Schmid and Gatica-Perez, Daniel — *Nov 2016*

- Training on the Job: Behavioral Analysis of Job Interviews in Hospitality. In: *Proc. of the 18th ACM on Int Conf. on Multimodal Interaction*, 2016

International Conference on Multimodal Interaction (ICMI) - 2016 — *Tokyo, Japan*

Finnerty, Ailbhe and **Muralidhar, Skanda** and Nguyen, Laurent Son and Pianesi, Fabio and Gatica-Perez, Daniel — *Nov 2016*

- Stressful First Impressions in Job Interviews. In: *Proc. of the 18th ACM on Int Conf. on Multimodal Interaction*, 2016

Mobile and Ubiquitous Multimedia (MUM) - 2016 *Rovaniemi, Finland*

Muralidhar, Skanda and Nguyen, Laurent Son and Costa, Jean and Gatica-Perez, Daniel *Dec 2016*

- Dites-Moi: Wearable Feedback on Conversational Behavior. In: *Proc. of the 15th Int Conf. on Mobile and Ubiquitous Multimedia*, 2016.

US Patent No. 9,426,405.23

Muralidhar, Skanda and Sathya, Vijay and Priyank Saxena *Aug. 2016*

- System and method of determining the appropriate mixing volume for an event sound corresponding to an impact related events and determining the enhanced event audio

Experience

Idiap Research Institute *Martigny, Switzerland*

Research Assistant (Prof. Daniel Gatica-Perez) *Dec. 2014 - Present*

- Currently investigating human nonverbal behavior at various workplace situations using multiple modalities (visual, audio and text), captured from ubiquitous and wearable devices.
- Developed an automatic, real-time feedback on user speaking behavior using Google Glass wearable device.

School of Computing, National University of Singapore *Singapore*

Graduate Student Researcher (Prof. Mohan Kankanhalli's lab) *Aug. 2012 - Jan. 2014*

- Investigated real-time human emotion detection using multiple modalities (audio and video).
- Explored use of the "Quantified Self" paradigm to classify human affect utilizing wrist-worn wearable.

AVK Systems S.A *Lausanne, Switzerland*

Senior Product Engineer *Feb 2011 to July 2012*

- Designed & optimized core modules of the proprietary slow-motion product.
- Actively involved in client interactions including presentations to FIFA, UEFA and the Swiss broadcaster (SRG/SSR).

Diginoc Technologies Pvt Ltd *Bangalore, India*

Product Engineer *Oct. 2009 to Jan. 2011*

- Led a **four** member team in developing key product modules on time.
- Wrote over 40% of the complete product (in **C**), currently evaluated at CHF 50 million.

Diginoc Technologies Pvt Ltd *Bangalore, India*

Research Associate *Oct. 2007 to Sept. 2009*

- Collected and analyzed over 2000 football kick sounds and their attributes leading to a patented technology critical to the company's product.
- Analyzed and developed algorithms (with Dr. Balaji Thoshkahna), for automatic detection and extraction of football kick sounds given a noisy stadium audio
- Developed algorithms to automatically (and in real-time) harmonize football sounds based on the background noise characteristics.

Dept of Electrical Engineering, Indian Institute of Science *Bangalore, India*

Undergraduate Thesis (Prof K. R. Ramakrishnan's lab) *July. 2004 to May. 2005*

- Developed algorithms to detect digital forgeries in images.
- Developed a new algorithm to reassemble fragmented colored images.

Skills

Programming	C/C++, Python, Matlab, R, UNIX shell scripting
Multimedia Packages	FFMPEG, OpenCV, SDL, OpenGL
Deep learning Frameworks	TensorFlow
Languages	English, Hindi, Sanskrit, French

Honors & Awards

2001	**University Scholarship,** Full scholarship during Bachelors	*Bangalore, India*
2005	**Best thesis award,** Dept of Electrical and Electronics Engineering, UVCE	*Bangalore, India*
2016	**Travel Grant,** ACM Int. Conf. on Multimodal Interaction	*Tokyo, Japan*
2017	**Travel Grant,** ACM Multimedia	*Mountain View, USA*

120

Extracurricular Activity

University Football Team — *Bangalore, India*

Member — *Sep 2002 - Sept 2004*

University Cricket Team — *Bangalore, India*

Member — *Sep 2002 - Sept 2004*

Black Belt in Karate — *Bangalore, India*

Martial artist — *Jan 2000*

121

**Ce document a été imprimé au Centre d'impression EPFL,
imprimerie climatiquement neutre, certifiée myClimate.**