Computational Analysis Of Behavior In Employment Interviews And Video Resumes

THÈSE N^{O} 6567 (2015)

PRÉSENTÉE LE 29 MAI 2015

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

LABORATOIRE DE L'IDIAP

PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Laurent Son NGUYEN

acceptée sur proposition du jury:

Prof. J.-Ph. Thiran, président du jury
Prof. D. Gatica-Perez, directeur de thèse
Prof. M. Chetouani, rapporteur
Prof. M. Schmid Mast, rapporteuse
Prof. S. Süsstrunk, rapporteuse

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2015

Acknowledgements

First of all, I would like to thank Prof. Daniel Gatica-Perez without whom this thesis would not have been possible. Thank you for the priceless support and guidance throughout this journey. Special thanks to the committee members, Prof. Jean-Philippe Thiran, Prof. Marianne Schmid Mast, Prof. Sabine Süsstrunk, and Prof. Mohamed Chetouani for taking the time to review this document, as well as for their valuable comments and feedback.

Thanks to the past and present collaborators of the SONVB and UBImpressed projects, Denise, Marianne, Daniel, Jean-Marc, Tanzeem, Kenneth, Ailbhe, Alvaro, Skanda, and Dayra. It was a real pleasure to work with all of you, and I feel lucky to be part of such an excellent team. This thesis was also greatly improved by the fruitful discussions with the Social Computing Group: Daniel, Oya, Dinesh, Minh-Tri, Joan, Dayra, Paco, Darshan, Gülçan, Rui, Skanda, Radu, Hari, Alvaro, Aleksandra, and Ailbhe. You guys are not only great to work with, but also really fun to hang out with. Thanks to Idiap Research Institute for being a great workplace, this truly makes a difference on a day-to-day basis. Special thanks to the Systems Group (especially Frank, Bastien, and Louis-Marie) for the technical support – with a smile. Without you guys, we would just be a bunch of kids trying to do research. Thanks to the admin team (Nadine and Sylvie) for the help. Thanks for the coffee drinkers for making these breaks fun (you know who you are). Thanks to the unihockey team for these epic games – and sorry to the ones I injured. Thanks to the babyfoot players for the awesome games. Thanks to the smoking crew for these cool fresh-air breaks.

I want to specially thank my wife Noémie for the unconditional support and love. We have gone a long way together and these last four years were super exciting. I am looking forward to the following ones, which I know will be great. Special thanks to my parents who have always supported me in my decisions. Thanks to the bunch of amazing friends that have always been here, be it for getting wasted, going to shows, or simply hanging out. Thanks to Tuco for being a great receptacle to blow off steam and a good excuse to have drinks during weekdays.

Abstract

Used in nearly every organization, employment interviews are a ubiquitous process where job applicants are evaluated by an employer for an open position. Consisting of an interpersonal interaction between at least one interviewer and a job applicant, they are used to assess interviewee knowledge, skills, abilities, and behavior in order to select the most suitable person for the job at hand. Because they require face-to-face interaction between at least two protagonists, they are inherently social, and all that recruiters have as a basis to forge their opinion is the applicant's behavior during the interview (in addition to his resume); in such settings, first impressions are known to play an important role. First impressions can be defined as snap judgments of others made based on a low amount of information. Interestingly, social psychology research has shown that humans are quite accurate at making inferences about others, even if the information is minimal.

Social psychologists long studied job interviews, with the aim of understanding the relationships between behavior, interview outcomes, and job performance. Until recently, psychology studies relied on the use of time-intensive manual annotations by human observers. However, the advent of inexpensive audio and video sensors in the last decade, in conjunction with improved perceptual processing methods, has enabled the automatic and accurate extraction of behavioral cues, facilitating the conduct of social psychology studies. The use of automatically extracted nonverbal cues in combination with machine learning inference techniques has led to promising computational methods for the automatic prediction of individual and group social variables such as personality, emergent leadership, or dominance.

In this thesis, we addressed the problem of automatically predicting hirability impressions from interview recordings by investigating three main aspects. First, we explored the use of state-of-the-art computational methods for the automatic extraction of nonverbal cues. As a rationale for selecting the behavioral features to be extracted, we reviewed the psychology literature for nonverbal cues which were shown to play a role in job interviews. While the main focus of this thesis is nonverbal behavior, we also investigated the use of verbal content and standard questionnaire outputs. Also, we did not limit ourselves to the use of existing techniques: we developed a multimodal nodding detection method based on previous findings in psychology stating that head gestures are conditioned on the speaking status of the person under analysis, and results showed that considering the speaking status improved the accuracy. Second, we investigated the use of supervised machine learning techniques for the prediction of hirability impressions in a regression task, and up to 36% of the variance could be explained, demonstrating that the automatic inference of hirability is a promising task. Finally, we

analyzed the predictive validity of thin slices, short segments of interaction, and showed that short excerpts of job interviews could be predictive of the outcome, with up to 34% of the variance explained by nonverbal behavior extracted from thin slices.
As another trend, online social media is changing the landscape of personnel recruitment. Until now, resumes were among the most widely used tools for the screening of job applicants. The advent of inexpensive sensors (webcams, microphones) combined with the success of online video hosting and viewing platforms (*e.g.*, YouTube) has enabled the introduction of a new type of resume, the video resume. Video resumes can be defined as short video messages where job applicants present themselves to potential employers. Video resumes hosted on online video sharing platforms represent an opportunity to study the formation of first impressions in an employment context at a scale never achieved before, and to our knowledge they have not been studied from a behavioral standpoint. We collected a dataset of 939 conversational English-speaking video resumes from YouTube. Annotations of demographics, skills, and first impressions were collected using the Amazon Mechanical Turk (MTurk) crowdsourcing platform. Basic demographics were then analyzed to understand the population using video resumes to find a job, and results showed that the population mainly consisted of young people looking for internship and junior positions. We conducted inference experiments to assess the amount of variance that could be explained by automatically extracted nonverbal cues, and results showed that most constructs could be inferred significantly more accurately than the baseline-average model, with up to 27% of the variance explained for extraversion, and up to 20% for social and communication skills.
We believe that our work is relevant for both organizational psychology and social computing. For psychologists, our study provides insights on what nonverbal cues might be used by recruiters to form the decision of hiring a person. Also, this thesis shows the feasibility of using automatically extracted cues to analyze nonverbal behavior in employment interviews and conversational video resumes, as an attractive alternative to manual annotations of behavioral cues. In social computing, our research has the potential to enable the development of several applications. For instance, the findings of this study could be used for the development of a training software application for job applicants by providing them with automatic feedback on simulated job interviews rehearsed at home. Another possible application would be the development of an online service to automatically screen job applicants, where candidates would be asked to provide, in addition to their resumes, a short video of themselves answering a series of predefined questions.

Keywords: social computing, face-to-face interactions, dyadic interactions, job interviews, first impressions, nonverbal behavior, hirability audio-visual feature extraction, online video resumes.

Résumé

Utilisés par la plupart des organisations pour la sélection de nouveaux collaborateurs, les entretiens d'embauches sont un processus dans lequel les candidats à un poste sont évalués par un recruteur. Les entretiens d'embauche sont constitués d'une interaction inter-personnelle entre au moins un recruteur ainsi que le candidat au poste et sont utilisés pour évaluer les connaissances, compétences, aptitudes et le comportement de celui-ci dans le but de sélectionner la personne la plus appropriée pour le poste. Étant donné qu'ils requièrent l'interaction en face-à-face entre au moins deux protagonistes, les entretiens d'embauche sont intrinsèquement sociaux ; ainsi, la seule base qu'ont les recruteurs pour prendre leur décision d'embauche est le comportement du candidat, ainsi que son *curriculum vitae* (CV). Dans ce type de situations, les premières impressions sont connues pour jouer un rôle important. Les premières impressions peuvent être définies comme des jugements instantanés basés sur une quantité minime d'informations. Il est intéressant de noter que la recherche en psychologie sociale a montré que les humains sont assez précis dans la formation de jugement sur les autres, même si l'information disponible est limitée.

La psychologie sociale a longuement étudié les entretiens d'embauche dans le but de comprendre les relations entre le comportement, les décisions d'embauche, et la performance au travail. Jusqu'à récemment, les études de psychologie se sont basées sur le codage manuel effectué par des observateurs humains. Or, l'émergence de caméras et de microphones bon marché, conjointement avec le développement de méthodes de traitement du signal élaborées, ont permis d'extraire des caractéristiques du comportement non-verbal de manière automatique et précise, facilitant la conduite d'études en psychologie sociale. L'utilisation de caractéristiques du comportement non-verbal obtenues de manière automatique en association avec des techniques d'apprentissage statistique ont conduit à des méthodes informatiques couronnées de succès pour la prédiction automatique de variables sociales telles que la personalité, l'émergence de leaders, ou la dominance.

Dans cette thèse, nous abordons le problème de la prédiction automatique d'impressions d'engageabilité à partir d'enregistrements d'entretiens d'embauche. Trois aspects principaux sont abordés. Premièrement, nous explorons la possibilité d'utiliser des méthodes de pointe pour l'extraction automatique de caractéristiques comportementales. Dans le but de sélectionner les comportements non-verbaux à extraire, nous avons examiné la littérature en psychologie et identifié les caractéritiques comportementales qui jouent un rôle lors d'entretiens d'embauche. Nous nous sommes focalisés principalement sur le comportement non-verbal, mais nous avons également étudié l'utilisation du comportement verbal ainsi que les résultats

de questionnaires psychométriques. Nous ne nous sommes pas limités à l'utilisation de méthodes existantes ; en effet, nous avons développé une méthode multimodale pour la détection de hochements de tête basée sur des résultats obtenus en psychologie qui démontrent que les mouvements de la tête sont conditionnés par l'état de parole de la personne analysée, et les résultats ont montré que la prise en compte de cet élément améliore la détection de hochements de tête. Deuxièmement, nous avons étudié l'utilisation de techniques d'apprentissage statistique supervisé dans le but de prédire les impressions d'engageabilité dans une tâche de régression, et nous avons réussi à expliquer jusqu'à 36% de la variance, démontrant que l'inférence automatique de l'engageabilité est une tâche prometteuse. Finalement, nous avons analysé la validité prédictive de tranche fines (*thin slices*) définies comme de courts segments d'interaction, et les résultats ont montré que de courts extraits de comportement pouvaient partiellement prédire la décision d'embauche basée sur l'entier de l'entretien d'embauche, avec des résultats atteignant 34% de variance expliquée.

En tant que nouvelle tendance, les médias sociaux sont en train de modifier le recrutement de nouveau personnel. Jusqu'à maintenant, les CVs ont toujours fait partie des outils les plus utilisés pour présélectionner les candidats à un poste. L'émergence de capteurs bon marchés (webcams, microphones) conjointement avec le succès de plateformes en ligne de visionnement de vidéos (par exemple, YouTube) a permis l'introduction d'un nouveau type de CV, le CV vidéo. Les CVs vidéo peuvent être définis comme de courts messages filmés où des demandeurs d'emploi se présentent à des employeurs potentiels. Les CVs vidéo hébergés sur les plateformes de partage de vidéos représentent une opportunité d'étudier la formation de premières impressions dans un contexte d'emploi à une échelle jamais atteinte jusqu'ici. À notre connaissance, les CVs vidéo n'ont pas été étudiés d'un point de vue comportemental. Nous avons collecté de YouTube une base de donnée de 939 CVs vidéo conversationnels en Anglais. Des annotations de données démographiques, de compétences, ainsi que de premières impressions ont été collectées en utilisant la plateforme de production participative (*crowdsourcing*) d'Amazon, Mechanical Turk (MTurk). Les données démographiques de base ont été ensuite analysées dans le but de caractériser la population de chercheurs d'emploi utilisant des CVs vidéo, et les résultats ont montré que ceux-ci sont principalement jeunes et recherchent des postes de stages ou de premier travail. Puis, nous avons effectué des expériences de prédiction afin d'évaluer la quantité de variance que peuvent expliquer des caractéristique non-verbales obtenues de manière automatique, avec des résultats atteignant 27% de variance expliquée pour l'extraversion, ainsi que 20% pour les compétences sociale et de communication.

Nous pensons que notre thèse est pertinente pour la psychologie du travail, ainsi que pour l'informatique sociale. Pour les psychologues du travail, notre étude fournit une connaissance sur les comportements non-verbaux pouvant avoir une influence sur les décisions d'embauche. Cette thèse démontre également la possibilité d'utiliser des caractérisiques comportementales extraites de manière automatique à partir d'entretiens d'embauche et de CVs vidéo en tant qu'alternative attrayante au codage manuel de comportements non-verbaux. En informatique sociale, notre recherche peut potentiellement ouvrir la possibilité de développer plusieurs applications. Par exemple, les résultats de cette étude pourraient être utilisés pour le déve-

loppement d'un programme informatique d'entraînement pour les chercheurs d'emplois en fournissant un feedback automatique sur leur comportement lors d'entretiens simulés réalisés dans un environnement sans risque. Une autre application possible pourrait être le développement d'un service en ligne pour pré-sélectionner des candidats à un poste, dans laquelle ceux-ci répondraient à une série de questions pré-enregistrées.

Mots-clés : informatique sociale, interactions en face-à-face, interactions dyadiques, entretiens d'embauche, premières impressions, comportement non-verbal, engageabilité, extraction audiovisuelle de caractéristiques, CVs vidéo.

Contents

List of Figures

List of Tables

1 Introduction

Used in nearly every organization, the employment interview is a ubiquitous process where job applicants are evaluated by an employer for an open position. The employment interview is an interpersonal interaction between one or more interviewers and a job applicant for the purpose of assessing the interviewee's knowledge, skills, abilities, and behavior in order to select the most suitable person for the job at hand, and is one of the most popular tools to perform this task [136]. Because they require face-to-face interaction between at least two protagonists, they are inherently social [68]. As applicants and interviewers meet for the first time, all that recruiters have as a basis to form their opinion is the applicant's verbal and nonverbal behavior during the interview (as well as their resumes), which makes this interaction very similar to what psychologists refer to as *zero-acquaintance situations* [18], interactions where protagonists are complete strangers.

In this context, first impressions play an important role. First impressions are snap judgments of others made on the basis of a low amount of information about the person [19]. Social psychology research has shown that the proverb "first impressions are the ones lasting" holds true up to a certain extent: humans are quite accurate at making inferences of others, even if the information is minimal [19]. Thus, minimal displays of behavior can be predictive not only of social constructs (*e.g.* personality or competence), but also of outcomes (*e.g.* teacher ratings) [19].

In face-to-face communication the spoken words form the verbal channel, while everything else represents nonverbal communication. Nonverbal behavior can be perceived aurally (through tone of voice, intonation, and amount of spoken time, for instance) and visually (through head gestures, body posture, gaze, or facial expressions) [78]. Interestingly, people quite often are able to perceive and interpret these social signals rapidly and correctly, and are the product of unconscious processes, which make them difficult to fake [78]. While the verbal channel remains the primary mode of communication, many social variables such as the judgment of personality, status, or competence (at the level of individuals), or the emergence of leadership or dominance (at the level of groups) are often outcomes of the multitude of micro-level nonverbal displays of behavior [78].

Nonverbal behavior in employment interviews has been studied by social psychologists for decades, mainly through the use of annotations of nonverbal cues by human observers. In the last decade, the advent of inexpensive audio and video sensors in conjunction with improved perceptual processing methods have enabled the automatic and accurate extraction of nonverbal cues, facilitating the implementation of social psychology studies [56, 134]. The use of automatically extracted nonverbal cues in combination with machine learning techniques has led to promising computational methods for the automatic inference of individual and group variables such as personality [29, 106], emergent leadership [119], or dominance [72]. To the best of our knowledge, employment interviews have not been previously analyzed using computational methods.

As another trend, online social media is changing the landscape of personnel recruitment. Beyond the massive success of LinkedIn and its 300+ million users from 200 countries [126], video interviews are beginning to modify the way in which applicants get hired. Until now, resumes were among the most widely used tools for the screening of job applicants in the personnel selection process [109]. The advent of inexpensive sensors (webcams, microphones) combined with the success of online video hosting and viewing platforms (*e.g.*, YouTube) has enabled the introduction of a new type of resume, the video resume. Video resumes can be defined as short video-recorded messages where job applicants present themselves to potential employers [65]. In comparison with traditional paper resumes, video resumes offer the possibility for applicants to show their potential, personality, and communication skills by displaying behavioral information visually and aurally [65], which makes video resumes similar to other forms of online video (*e.g.*, video blogs) in terms of setting [29]. At this present day, video resumes are an emerging phenomenon. Related work on video resumes is scarce and has focused on their reception by recruiters [65, 77]. To our knowledge, no study has investigated video resumes from a behavioral standpoint. Video resumes hosted on online video sharing platforms represent a unique opportunity to study the formation of first impressions in an employment context at a scale never achieved before.

1.1 Objectives

The aim of this thesis is to design and develop a computational framework for the analysis of first impressions in the personnel selection process. From this broad objective, several research questions were posed. In employment interviews, can first impressions of hirability be inferred automatically? What cues are predictive of hirability impressions? Is language style predictive of interview outcomes? Can short excerpts of job interviews be predictive of hirability? Can first impressions from video resumes be rated consistently? Can they be inferred using fully automated methods?

To our knowledge, these questions have not yet been studied from a computational standpoint. Addressing them required a multidisciplinary approach: computer vision, audio processing, and machine learning techniques were necessary for the extraction of behavioral features and

the inference of social variables, while the decades of related work in social psychology were essential for selecting behavioral cues to be extracted and interpreting results appropriately.

1.2 Motivation

We believe that our work is relevant for both organizational psychology and social computing. For psychologists, our work provides insights on what nonverbal cues might be used by recruiters to make the decision of hiring a person during an employment interview. Also, our work shows the feasibility of using automatically extracted cues to analyze nonverbal behavior in employment interviews, as an attractive alternative to manual annotations of behavioral cues. In social computing, our research has the potential to enable several applications. For instance, the findings of this thesis could be used for the development of a training software application for job applicants by providing them with automatic feedback on simulated job interviews rehearsed at home. Another possible application would be the development of an online service to automatically screen job applicants, where candidates would be asked to provide, in addition to their resumes, a short video of themselves answering a series of predefined questions.

1.3 Summary of contributions

The contributions of this thesis are the following:

1. **Collection of employment interviews**. To investigate the feasibility of building a computational framework for the automatic inference of interview outcomes, and due to the lack of a publicly available job interview corpus, we designed the technical infrastructure and collected a dataset of 62 real job interviews, where participants were applying for a marketing job. The interviews were dyadic, and audio and video were recorded for both the applicant and the interviewer. Social variables were collected through the use of questionnaires and annotations.

2. **Multimodal detection of head nods**. Head nods are vertical up-and-down movements of the head, rhythmically raised and lowered, and are used in virtually every face-to-face interaction to signal a *yes*, display interest, or anticipate an attempt to capture the floor [16]. Social psychology has established the relationship between applicant head nods and job interview outcomes [57]. While the value of using context from the perspective of the speaker to improve the detection of head nods of the listener has been established, one aspect that has not been studied in detail is the effect of the conversational self-context on head nod detection. To detect head nods in natural interactions, we developed a multimodal method that leverages a finding in psychology that states that head gestures are conditioned by the speaking status of the person of interest [62]. We showed that using the audio self-context improved the head nod detection accuracy.

3. **Extraction of behavioral cues**. We used state-of-the-art audio and video processing methods to extract behavioral cues from the interview recordings. As rationale for selecting the behavioral features to be extracted, we studied the psychology literature and identified nonverbal cues that were shown to play a role in job interviews. Three types of behavioral cues were extracted: (1) Nonverbal cues were automatically extracted from the audio modality (turn- and prosody-based cues) and the video modality (head nods, head motion, overall motion), and multimodal and relational cues were built by combining modalities and persons of interest (*e.g.*, mutual nods, nodding while speaking). (2) Body communication cues from the applicant were obtained from manual annotations of postures and gestures and automatically extracted descriptors derived from hand velocity. (3) Linguistic and paralinguistic features were extracted from manual transcriptions of the applicant's speech, using the Linguistic Inquiry Word Count (LIWC) system [130].

4. **Inference**. We investigated the use of behavioral features for the automatic inference of hirability impressions. To understand the relationship between nonverbal cues and hirability, we conducted a correlation analysis. The inference task was defined as a regression task, and several regression and dimensionality reduction methods were tested. To understand what cues were used by raters while forming their hirability impressions, the predictive validity of feature groups (defined by modality and person of interest) were analyzed. We analyzed the use of psychometric questionnaires widely used in the personnel selection process for the prediction of hirability scores. We assessed the use of body communication cues for the prediction of personality traits and hirability impressions. Last, we analyzed the predictive validity of verbal categories extracted from manual transcriptions. Results showed that up to 36% of the variance could be explained using nonverbal cues as predictors, demonstrating that the automatic inference of hirability impressions was a feasible task. Questionnaire data and verbal behavior were not predictive of hirability and did not increase the accuracy when combined with nonverbal behavior, suggesting that hirability impressions were formed on the basis of nonverbal behavior.

5. **Thin slices of employment interviews**. We analyzed the predictive validity of thin slices, *i.e.* short excerpts of employment interviews. In order to account for the structured nature of the job interviews, slices were defined by the questions posed during the interview. Audio-visual nonverbal cues were extracted from these thin slices and used for the inference of hirability impressions gathered from the full interaction. Although behavioral cues extracted from thin slices were not as accurate as the full interaction, they were still predictive of the interview outcome. In comparison with the observer predictive validity, *i.e.* the level of agreement between ratings obtained from thin slices and the full interaction, results obtained from nonverbal cues extracted from thin slices were competitive. No slice stood out in terms of predictive validity: all slices yielded comparable results, suggesting that the observed nonverbal behavior did not drastically change from one slice to another.

6. **Analysis of online video resumes**. Online conversational video resumes represent a unique opportunity to study first impressions related to personnel selection at a scale never achieved before. Presently, little is known about the demographics of video resumes or the type of social constructs that can be inferred from them. To fill this void, we collected a dataset of 939 English-speaking conversational video resumes from YouTube. Analysis of the demographics of the dataset showed that job-seekers using video resumes were mainly young professionals looking for internship or junior positions, and that two thirds of the population were men. Annotations of employment-related social constructs (skills, personality, hirability) were collected by naïve raters using Amazon Mechanical Turk, and most variables were found to be reliable upon analysis of the interrater agreement, not only suggesting that raters competed the task conscientiously, but also that the annotations of first impressions on video resumes by unacquainted judges was a feasible task. To understand the structure underlying the perceived skills, we conducted a data-driven clustering analysis and results showed that skills could be grouped into three high-level clusters, namely professional skills, communication skills, and social skills. Nonverbal cues were automatically extracted, and we evaluated several regression methods for the inference of personality and hirability. Results demonstrated the feasibility of automatically inferring hirability, extraversion, and openness to experience in video resumes to some degree, achieving R^2 results up to 27%, thus confirming our initial hypothesis that first impressions on video resumes were at least partly based on nonverbal behavior.

1.4 Outline

A graphical summary of this thesis is displayed in Figure 1.1. In Chapter 2, we discuss the related work in social psychology and social computing. In Chapter 3, we present the SONVB[1] job interview corpus, a dataset of real employment interviews that was used for the main part of this thesis. In Chapter 4, we discuss the methods used to extract behavioral cues from the job interview recordings. In Chapter 5, we propose a computational framework for the automatic inference of interview outcomes and analyze in detail the role of behavioral cues in the formation of hirability impressions. In Chapter 6, we investigate the use of short excerpts of interviews for the inference of hirability impressions. In Chapter 7, we present the collection and analysis of conversational video resumes obtained from YouTube. In Chapter 8, we conclude this thesis and discuss the limitations of our work as well as possible avenues for future work. Please note that in order to facilitate the readability of this dissertation in disciplines other than computer science, we keep the number of equations included in the document to a minimum.

[1] SONVB is the acronym used for Sensing Organizational NonVerbal Behavior, the name of the project.

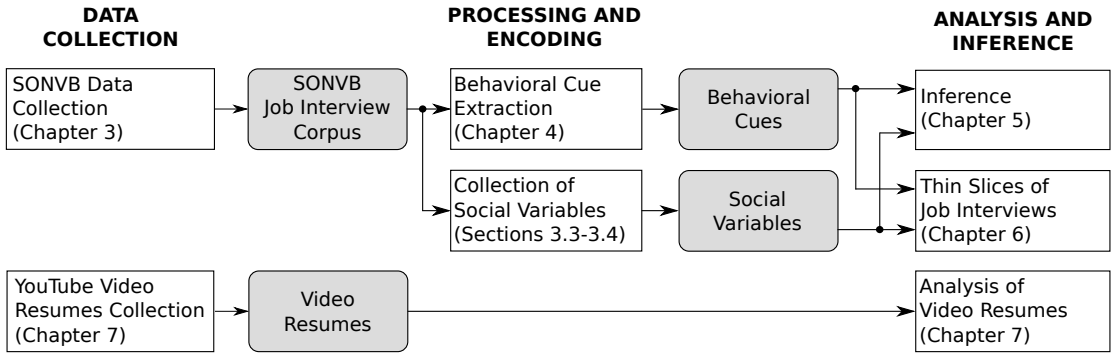


Figure 1.1 – Overview of the thesis.

1.5 Publications

This thesis compiles the work published in peer-reviewed journal and conference papers listed in Section 1.5.1, in addition to unpublished work. The candidate also collaborated on other papers, which are listed in Section 1.5.2 but not discussed in this thesis.

1.5.1 Publications as first author

L. S. Nguyen, J.-M. Odobez, D. Gatica-Perez. "Using Self-Context for Multimodal Detection of Head Nods in Face-to-Face Interactions", in *Proceedings of the ACM international Conference on Multimodal Interaction (ICMI)*, 2012. Conference paper.

L. S. Nguyen, A. Marcos-Ramiro, M. Marron-Romera, D. Gatica-Perez, "Multimodal Analysis of Body Communication Cues in Employment Interviews", in *Proceedings of the ACM international Conference on Multimodal Interaction (ICMI)*, 2013. Conference paper.

L. S. Nguyen, D. Frauendorfer, M. Schmid Mast, D. Gatica-Perez, "Hire Me: Computational Inference of Hirability in Employment Interviews Based on Nonverbal Behavior", in *IEEE Transactions on Multimedia*, 16(4): 1018-1031, 2014. Journal article. Idiap Student Paper Award 2014.

L. S. Nguyen, D. Gatica-Perez, "Hirability in the Wild: Analysis of Online Conversational Video Resumes", submitted for journal publication.

1.5.2 Publications as coauthor

D. B. Jayagopi, S. Sheikhi, D. Klotz, J. Wienke, J.-M. Odobez, S. Wrede, V. Khalidov, **L. S. Nguyen**, B. Wrede and D. Gatica-Perez, "The Vernissage Corpus: A Conversational Human-Robot-Interaction Dataset, in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2013. Conference paper.

A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, **L. S. Nguyen** and D. Gatica-Perez,

"Body Communicative Cue Extraction for Conversational Analysis", in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2013. Conference paper.

K. A. Funes Mora, **L. S. Nguyen**, D. Gatica-Perez, J.-M. Odobez, "A Semi-Automated System for Accurate Gaze Coding in Natural Dyadic Interactions", in *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2013. Conference paper.

D. Frauendorfer, M. Schmid Mast, **L. S. Nguyen**, D. Gatica-Perez, "Nonverbal Social Sensing in Action: Unobtrusive Recording and Extracting of Nonverbal Behavior in Social Interactions Illustrated with a Research Example". *Journal of Nonverbal Behavior. Special Issue: Contemporary Perspectives in Nonverbal Research*, 38(2):231-245, 2014. Journal article.

N. A. Murphy, J. A. Hall, M. Schmid Mast, M. A. Ruben, D. Frauendorfer, D. Blanch-Hartigan, D. L. Roter, **L. S. Nguyen**, "Reliability and Validity of Nonverbal Thin Slices in Social Interactions", in *Personality and Social Psychology Bulletin*, in press. Journal article.

M. Schmid Mast, D. Gatica-Perez, D. Frauendorfer, **L. S. Nguyen**, T. Choudhury, "Social Sensing for Psychology: Automated Interpersonal Behavior Assessment", in *Current Directions in Psychological Science*, in press. Journal article.

2 Related Work

In this chapter, we review the literature on different aspects in social psychology, social computing, and multimedia analysis that relates to our work, *i.e.* the computational analysis of employment interviews and video resumes. Understanding the basis on which hiring decisions are made is a key research area in social and organizational psychology. The automatic inference of social constructs is relevant in social computing, where high-level social variables such as personality, dominance, or emergent leadership have been analyzed from recordings of face-to-face interactions. To our knowledge, hirability impressions in employment interviews have to this day not been analyzed from a computational standpoint, and our work contributes to the literature by filling this gap.

This chapter is structured as follows. In Section 2.1, we define nonverbal communication and discuss its communicative functions and properties. In Section 2.2, we review the existing computational methods to automatically extract nonverbal cues. In Section 2.3, we review the literature related to verbal behavior. In Section 2.4, we discuss the related work in social psychology on employment interviews. In Section 2.5, we discuss the related work on the computational methods previously used to infer social constructs from face-to-face interactions, as well as the computational literature related to employment interviews and similar settings. In Section 2.6, we review the literature related to video resumes. We finally conclude this chapter in Section 2.7.

2.1 Nonverbal communication

In face-to-face communication the spoken words form the verbal channel, while everything else represents nonverbal communication. Nonverbal behavior can be perceived aurally (through tone of voice, intonation, and amount of spoken time, for instance) and visually (through head gestures, body posture, gaze or facial expressions) [78]. Interestingly, people quite often are able to perceive and interpret these social signals rapidly and correctly, and are the product of unconscious processes, which make them difficult to fake [78]. While the verbal channel remains the primary mode of communication, many social variables such as the

judgment of personality, status, or competence (at the level of individuals), or the emergence of leadership or dominance (at the level of groups) are often outcomes of a multitude of micro-level nonverbal displays of behavior [78]. Nonverbal behavior comprises four main categories, which are described below:

1. **Vocalics** (*a.k.a.* acoustics) relate to the way through which a vocal message is conveyed, in opposition to the content of the message. In other words, acoustics refer to *how* something is said rather than *what* is being said. Vocalic nonverbal behavior can be expressed through prosody, turn-taking patterns, linguistic and non-linguistic vocalizations, and pauses [134]. Prosody is the term designed to describe all the variations in the vocal delivery that accompany speech, and is composed of three fundamental acoustic properties, namely speech rate (the number of voiced segments[1] per unit of time), pitch (the fundamental frequency), and intensity (the energy of the speech sound, perceived as loudness) [78]. Turn-taking is another important aspect of vocalics and relates to the regulation of conversations and the coordination during the speaker transitions [134]. Linguistic vocalizations refer to all the non-words used as if they were words (*e.g.*, *"ehm"*, *"ah-hun"*), and are used to replace words that for some reasons cannot be found in embarrassing or difficult situations; linguistic vocalizations can also be used to produce back-channels, short vocalizations emitted in conjunction during listening turns to acknowledge understanding or agreement. Non-linguistic vocalizations are vocal outbursts that include laughing, crying, whispering, or sighing. Silences are defined as non-speech segments, and can occur when a person has difficulties talking (hesitation silence), needs some time to think about what was or will be said (psycholinguistic silence), or as a way to convey a message about the interaction taking place (interaction silence). In interactions, vocal cues play an important role in managing the interaction; they are used to yield, request, maintain, or deny the floor. Vocal cues can reveal a wide array of personal social constructs such as personality traits, sociodemographic characteristics, age, status, as well as emotional states [78]. During employment interviews, vocal cues are constantly expressed through pitch, speech rate, vocal energy, turn-taking, pauses, vocalizations, and back-channels.

2. **Kinesics** relate to the nonverbal behavior related to body movement of any part of the body or the body as a whole. It includes facial expressions, gaze, hand gestures, posture, head gestures. Facial expressions are used to communicate affective states, such as anger, disgust, pain, happiness, or fear [78]. Eye gaze refers to where a person is looking at, and is used to regulate the flow of communication, monitor feedback, reflect cognitive activity, express emotions, and communicate the nature of the interpersonal relationship [78]. Hand gestures are an essential component of kinesics as they are used to enrich the vocal content and aid listener comprehension by augmenting the attention, activating images or representations in the listener's mind, and increasing the recall of what is being said [78]. Body posture is another important component

[1] Voiced regions refer to segments of the articulatory process in which the vocal cords vibrate, and roughly correspond to vowels [24].

of kinesics as various emotions such as fear, sadness, or happiness can be inferred from a person's posture; in conversations, body postures can be used as markers: for instance, changes of body posture can precede a long utterance and may be kept for the duration of the speaking turn [78]. Head nods are used *inter alia* to signal a 'yes', display interest, enhance communicative attention by occurring in synchrony with the other's speech, or anticipate an attempt to capture the floor [62, 16]; head nods play a role in back-channeling, can reveal personal characteristics, and even predict social outcomes [57].

3. **Proxemics** relate to the interpersonal distance between protagonists during face-to-face interactions. It also includes the arrangement of space where the interaction occurs, such as the distribution of furnitures or the overall space allowed for each interaction protagonist, and has shown to affect the interaction: for instance, sitting side-by-side can elicit collaborative interactions, while opposite settings can yield competitive behavior. Interestingly, physical environments can tell a lot about their owners: offices and bedrooms were found to be predictive of their owners' personality traits [61]. During interactions, co-occurring proxemic shifts forward or backward can mark important segments of the conversations, such as beginnings, ends, or topic changes [78]. In job interviews, the distribution of space is fixed across job applicants, and the protagonists usually face each other at two sides of a table. The only behavioral cue of this category which can be displayed is therefore the interpersonal distance between the applicant and the interviewer, defined by the body posture (leaning forward / leaning backward). In video resumes, proxemics can be expressed *via* the distance between the applicant and the camera.

4. **Haptics** refer to the way humans communicate and interact through the use of the touching sense. Haptics can be self-focused or expressed towards others. The communicative functions of interpersonal touch are highly dependent on the context and the relationship between the protagonists. In professional environments, handshakes are used as part of welcoming greetings and can convey warmth, or friendliness. In close relationships, haptics are used to express emotions, appreciation, sexual attraction, and support [78]. Touch can also be associated with negative affects such as anger and frustration and can be expressed by hitting and slapping, although such behaviors are more likely to happen among young children than adults. Some self-touch behaviors are referred to as *adaptors* and are behavioral adaptations in response to certain situations; they are generally associated with negative feelings, such as psychological discomfort, anxiety, guilt, stress, or suspiciousness [78]. During employment interviews (apart from the welcoming handshake), behaviors related to haptics are generally self-focused.

2.2 Automated extraction of behavioral cues

In the last decade, the advent of inexpensive audio and video sensors in conjunction with improved perceptual processing methods have enabled the automatic and accurate extraction

of a number of nonverbal cues. Because it is expressed both visually and aurally, nonverbal behavior is inherently multimodal, and the related work aiming at its automatic extraction spans both the audio processing and computer vision literatures. In this section, we review the literature on automatic extraction of nonverbal behavioral cues. Because of the large number of existing methods to extract nonverbal cues, we do not intend to be exhaustive in the review; rather, we present basic challenges and examples of recent methods.

2.2.1 Audio cues

The process of automatically detecting the person speaking during an interaction is known as speaker diarization, and consists of three main steps: the first is the classification of the audio stream into speech and non-speech segments; the second is the detection of speaker transitions and is referred to as speaker segmentation; the third is speaker clustering and its purpose is to group speech segments based on speaker characteristics [79, 134]. In meetings, the diarization problem can be simplified through the use of microphone arrays that can localize sound sources *via* acoustic beamforming [21]. The Microcone device [6] is a commercial product based on this concept.

Prosodic features (voiced rate, pitch, and energy) are generally extracted on the basis of a two-step hidden Markov model (HMM) that is used to segment the audio signal into speech/non-speech and voiced/unvoiced regions [24]. The number of voiced segments per unit of time provide an indication of the speaking rate of the person under analysis. Then, for each voiced region, pitch is tracked using a probabilistic method [24]. The energy is a property of any digital signal and can be obtained by summing the squares of each sample [134]. Several software packages to extract prosodic features are publicly available, such as Praat [9], Wavesurfer [12], or the MIT Media Lab speech feature extraction code [7].

2.2.2 Visual cues

Detecting faces robustly and accurately constitutes a strict requirement for the problem of automatically extracting kinesic cues related to facial expressions, gaze, and head nods. Numerous methods for the automated detection and tracking[2] of faces exist, which are based on geometric features (*e.g.* shapes of the mouth, eyes, etc.), location of facial points (*e.g.* corners of the eyes, mouth, nose, etc.), face texture, or skin color [134]. The size of the face bounding box can be used as a proxy for interpersonal distance when the camera-person relative placement is known [29].

Most methods to automatically detect head nods have been developed in the context of human-computer interfaces (HCI). The primary goal of these studies was to enable a machine to detect a 'yes' signaled by a head nod. Within this context, some studies proposed to track in-

[2] Face detection refers to the problem of detecting one or more faces in an image, whereas face tracking aims at following faces *e.g.* for the duration of a video, and includes temporal constraints in the model.

terest points of the face [75, 129] and use state-based approaches such as finite state machines [47] and hidden Markov models [96, 129] to detect nodding. These approaches showed good performance in restricted contexts where the head motions were explicit; however, they did not allow to detect subtle head movements that often occur in natural face-to-face interaction. Communicative contextual audio-based features have been used to improve the detection of head nods in dyadic scenarios [96, 94], but the interactions used for experiments were not entirely natural. While the value of using audio-based context (from the perspective of the speaker) to improve the detection of listener head nods has been established, one aspect that to our knowledge has not been studied in detail is the effect of the audio-based self-context on head nod detection (*i.e.*, the speaking status of self).

Recently, there have been important advances in computer vision to develop automatic gaze tracking methods [64]. The first studies focused on human-computer interface (HCI) applications in constrained settings, such as people looking at computer screens. Most early systems were either invasive, expensive, or required constrained body and head movement. Less restrictive systems for the analysis of natural interactions, capable of estimating head and gaze information, only managed to provide coarse measurements, such as gaze pan [58]. Recent studies in gaze estimation have aimed at overcoming these limitations [54, 85]. Provided RGB-D data (*i.e.*, depth information in addition to standard color images), methods have been proposed to remotely extract head pose and gaze, which yields competitive results under unrestricted head movements and low resolution [54]. Despite the recent advances of gaze estimation methods in unconstrained settings, it still remains a challenging problem in natural interactions [54].

The automatic recognition and analysis of facial expressions has been an active topic in computer science for over two decades [133]. One of the main focus of this research topic is the recognition of action units composing the Facial Action Coding System (FACS) [39]. In FACS, facial expressions are decomposed into 46 action units, which correspond to the activation of specific muscles of the face [39], and all possible facial expressions can be seen as a combination of action units. It is commonly accepted from the psychological theory that human emotions can be classified into six categories, namely surprise, fear, disgust, anger, happiness, and sadness [84], and these archetypal emotions can be derived from the combination of specific action units [39]. A large amount of related work has focused on constrained settings, where the six basic emotions were posed, which can strongly differ from natural interactions where the display of affective states are often subtle. In the last decade, several studies have emerged on the automatic analysis of spontaneous facial expressions [23, 40]. Recently, software packages for automatically extracting facial expressions in real-time have appeared, such as the Computer Expression Recognition Toolbox (CERT) [84], which enabled the conduct of affective analyses at a large scale [30]. However, analyzing facial expressions during natural interactions where protagonists are speaking remains a challenge due to the noise generated by the head and mouth movements, as well as the subtle displays of facial expressions.

Automatic markerless motion capture has recently made huge progress: body regions were accurately classified from RGB-D images [124] resulting in the commercialization of the highly successful Microsoft Kinect device. However, this method was designed for gaming applications where the full body is visible, the movements are broad, and the users adapt their behavior to the system. In contrast, Kinect can fail to correctly register body posture in conversational settings where people are seated and display natural movements. A method has used motion energy images to summarize a video into one image where each pixel intensity corresponds to the motion displayed throughout the video at this location, providing a coarse estimation of body activity [29]. In another study, body motion has been detected in group interactions from accelerometer sensor data from wearable devices [51]. Recently, a method was proposed to automatically extract hand location, speed, as well as body pose from standard video recordings of seated people in a conversational setting [88].

2.3 Verbal content

The words we use while speaking or writing reveal many aspects of our identity [105, 38]. Although previous work in psychology has shown that language style is associated with various personal social constructs, such as personality, affective states, status, deception, demographics (gender, age), or culture [38], little is known about the role of verbal content in the formation of hirability impressions. Among the few studies investigating the role of language style in the employee selection process, nonverbal behavior conditioned by the level of verbal content predicted high ratings in simulated job interviews [112].

Recent studies in social computing have investigated the use of verbal content for the prediction of social constructs. Although the first studies were interested in the prediction of constructs based on written text from blogs (*e.g.*, [101]), recent studies have also investigated natural spoken language style for social construct prediction. For instance, emergent leaders in small groups [121], or personality impressions of YouTube video bloggers [31] were predicted from manually annotated transcriptions and automatically extracted representations of verbal behavior. Related to personnel selection, personality traits were predicted from blog posts of LinkedIn users on an e-recruitment platform, using linguistic features as predictors [49].

Several representations of verbal content have been successfully used for the prediction of social variables. Widely used in natural language processing, n-grams are a data-driven representation of verbal content which encode the occurrence of n subsequent words, and have been successfully used for the inference of Big-Five personality traits (*e.g.*, [70]). Due to the high dimensionality of the obtained feature vector, n-grams are usually better used for large datasets. As a high-level representation of verbal content, Linguistic Inquiry Word Count (LIWC) is a computer software resulting from years of social psychology research focused on the validation of the psychometric properties of a word categorization system that relates linguistic and paralinguistic categories to psychological constructs [130]. In its original English

version, LIWC is built based on a dictionary of 4500 words and word stems. Each word found in the dictionary is assigned to one or several categories, corresponding to linguistic and psychological processes, personal concerns, spoken categories, and punctuation. Several computational studies have successfully used LIWC as a representation of verbal content for the prediction of constructs from written text [36] or natural spoken language [121, 31]. However, to the best of our knowledge no computational study has investigated the role of language use in job interviews.

2.4 Employment interviews and nonverbal behavior in psychology

For decades, employment interviews have been a major research topic in social psychology. In particular, previous works have investigated the reliability (*i.e.*, the level of agreement between judges for rating applicants) [69] and validity (*i.e.*, the amount of correlation between interview ratings and job performance) [110] of employment interviews, as well as the relationship between high-level social variables (*e.g.* personality traits, general intelligence) and job performance [117, 22]. Particular attention has been put on the impact of the applicant's nonverbal behavior on the job interview outcome. For instance, Imada and Hakel [71] showed that applicants who use more immediacy nonverbal behavior (*i.e.*, eye contact, smiling, body orientation toward interviewer, less personal distance) were perceived as being more hirable, more competent, more motivated, and more successful than applicants who did not. Forbes and Jackson [53] showed that applicants who were employed made more direct eye contact, smiled more, and nodded more during the job interview than applicants who were rejected. Parsons and Liden [104] found that speech patterns explained a remarkable amount of variance in the hiring decision, beyond and above objective information. Also, Anderson and Shackleton [20] reported that applicants who were selected made more eye contact and produced more facial expressions during the job interview than non-accepted applicants. One explanation for the positive relation between applicant nonverbal behavior and hiring decision can be based on the immediacy hypothesis, which establishes that the applicant reveals through his or her immediacy behavior (eye contact, smiling, hand gestures, etc.) a greater perceptual availability, which leads to a positive effect on the interviewer and therefore to a favorable evaluation [71]. In these studies, all coding of nonverbal behavior was done manually. Also, these works were not addressed as a prediction task in the machine learning sense (*i.e.*, no separation between training and test data was done) and the analyses rather focused on correlation and in-sample ordinary least-squares linear regression.

Our work contributes to the organizational psychology community by proposing an automated framework to analyze employment interviews. We believe that this framework has the potential to streamline the conduct of social psychology studies by reducing the need for time-intensive manual annotations of nonverbal behaviors.

2.5 Employment interviews in social computing

In the last decade, numerous studies have investigated the use of computational approaches for the analysis of interactions from the perspective of nonverbal behavior. These automated frameworks have been used for the prediction of interest [137], dominance [72], emergent leadership [119], roles [50], and personality traits [106, 25, 29] from sensor data in small groups. Other studies have also examined dyads, mainly for the prediction of outcomes in speed-dating [87] or negotiations [44, 103], but also to identify indicators of psychological disorders [122].

Most existing methods for automatic inference of social constructs consist of two main steps. In the first step, behavioral features are extracted from audio (turn-taking, prosody, *e.g.* [106, 119, 72]) and video (body and head activity, visual focus of attention, *e.g.* [103]). In the second step, machine learning algorithms (including hidden Markov models [128], probabilistic graphical models [128], support vector machines [29, 25], or topic models [73]) are trained and used to infer the social constructs at hand.

In the context of organizations, Curhan and Pentland investigated the relationship between automatically extracted audio nonverbal cues and the outcome of simulated dyadic job negotiations [44]. To this end, they designed a negotiation scenario between a vice president and a middle manager who had to agree on an employment package. The outcome of each negotiation item would earn points for each party, but the scoring system was designed such that the goals of the protagonists were often conflicting. Audio nonverbal cues (activity, engagement, emphasis, mirroring) were automatically extracted for both protagonists from five-minute excerpts of the interaction and were used to predict individual and joint negotiation scores. For individual scores, the authors reported results of $R^2 = .30$ using an ordinary least-squares linear regression model with no cross-validation.

Related to employment interviews, Batrinca *et al.* [25] used a computational approach to infer Big-Five personality traits in self-presentations, where 89 participants had to shortly introduce themselves in front of a computer in a setting similar to video resumes or video interviews. A mixture of manually- and automatically-obtained audio and video nonverbal cues (gaze, frowning, hand movement, head orientation, posture, prosody) were extracted from the self-presentation videos and used to predict Big-Five personality traits in a classification task, where binary low/high classes were balanced. To this end, support vector machines with greedy feature selection were used, and classification accuracies above 0.7 were obtained for the traits of conscientiousness, emotional stability, and extraversion. The authors assumed a close link between the constructs of personality and hirability; they however did not explicitly address the problem of automatic hirability prediction.

Within the context of human-computer interfaces (HCI), Baur *et al.* [26] recently proposed a system making use of a virtual character and signal processing techniques to create an employment interview simulation virtual environment. In this immersive environment, the virtual agent played the role of a recruiter and reacted and adapted to the applicant's behavior.

To this end, hand-to-face, looking away, postures, leaning forward/backward, voice activity, and smiles were automatically extracted from audio and RGB-D data recorded from Kinect. The overall system was qualitatively evaluated by the six participants who tested the system.

Along the same lines, Ehsan *et al.* [67] developed a system to train social skills based on a virtual agent that read facial expressions, speech, and prosody, and responded with verbal and nonverbal behavior in real time. The system was used in the context of a training for job interviews, where participants interacted with the system. Participants received two types of feedback: first, they were shown their videos including visualizations of smiles, head shakes, head nods, and speech intensity; second, they were provided a summary feedback consisting of their displayed behavior, including smiles, speaking rate, use of weak language, response time, and pitch variation. Participants who interacted with the system and received both feedbacks earned higher expert ratings in a post-training assessment compared to participants who did not interact with the system or who only received video feedback.

Very recently, Chen *et al.* [37] proposed a multimodal method to recognize job applicants' affective states from video interviews. To this end, they collected a dataset of 20 video interviews, where four amateur actors were instructed to act three levels of composure (bad/medium/good), as well as exaggerated versions of bad/good composures. The six universal emotions (surprise, fear, disgust, anger, happiness, and sadness), as well as disgust, were inferred from prosodic features and facial expressions extracted using the Computer Expression Recognition Toolbox (CERT) [84], followed by a greedy feature selection step.

To our knowledge, our work is the first explicitly addressing the issue of inferring hirability in employment interviews. Our work approaches this problem from a face-to-face, behavioral perspective where sensing, feature extraction, and social inference are automated. Furthermore, we make use of interviewer, applicant, and relational behavioral cues extracted from both the audio and visual modalities as predictors for the regression task of inferring expert-coded hirability scores.

2.6 Video resumes

Despite the emergence of video resumes and video interviewing platforms as a new medium for personnel recruitment, scholarly publications on this topic are still scarce. The first references to video resumes in the scientific community date from 1992, when Kelly *et al.* [76] proposed to use video resumes as a tool to help deaf college students develop communication skills to help them secure a job position. One year later, Rolls *et al.* proposed the use of video resumes as a way to "supply the potential employer with insight into the student's personality and character" [116]. In this conceptual study, the authors described the method to record these early versions of video resumes: they lasted around 15 minutes, were highly structured, and consisted of behavioral questions (*i.e.*, questions related to past experiences in specific situations). However, no analysis on the type of first impressions that were made on potential recruiters was completed.

Recently, a doctoral thesis by Hiemstra [65] examined the use of video resumes in the personnel selection process. The main focus of this work was to investigate the fairness and discriminatory effects of paper and video resumes. To this end, 445 job-seekers were enrolled in a 2-day training program organized by the Dutch government, where they received training and recorded their video resume. A professional studio then edited each recording, resulting in a personal 40-60 second video resume for each participant. This study did not investigate the role of nonverbal behavior in the formation of first impressions.

Kemp *et al.* [77] examined the perception of video resumes by sales recruiters. Their study focused on identifying general perceptions of video resumes among recruiters and their reactions to video resumes as a screening tool. To this end, video resumes of 10 students were recorded professionally and shown to a pool of recruiters, who were instructed to complete a questionnaire related to their perceptions of video resumes. In this work, the perception of how the person behaved and performed was not investigated.

Within the field of social video analysis, Biel *et al.* [29] have investigated the formation of personality impressions in conversational video blogs. To this end, a dataset of 442 conversational video blogs was collected from YouTube, and personality impressions were annotated by naïve judges on the Amazon Mechanical Turk crowdsourcing platform. To understand the basis on which personality impressions were made, both nonverbal [29] and verbal [31] behavioral cues were automatically extracted from the videos and used to infer personality impressions. However, the studied data was not related to job search as video resumes.

To the best of our knowledge, conversational video resumes have not been analyzed from a computational, nonverbal perspective. We believe that online video resumes constitute an interesting setting for the study of the formation of first impressions, and our work aims at filling this gap.

2.7 Conclusion

In this chapter, we reviewed the related work on employment interviews and video resumes. We first presented the taxonomy and communicative functions of nonverbal behavior, and discussed the state-of-the-art methods to automatically extract nonverbal cues, which will serve as a basis for obtaining an accurate representation of the displayed behavior of protagonists in employment interviews. We then discussed the use of verbal behavior for the inference of social variables, as well as its possible representations. Employment interviews have been studied for decades by psychology researchers, and several relationships between nonverbal behavior and hirability impressions have been established. This important body of related work is useful as it provides us with hints on what nonverbal cues are associated with hiring decisions, therefore it serves as a catalog for the behavioral cues to be extracted. In social computing, relatively few studies have investigated employment interviews, despite their ubiquity in the personnel selection process. To our knowledge, our work constitutes the first aiming at inferring hirability impressions automatically. The related work in social computing focusing

on the inference of social constructs in face-to-face interactions is however sufficiently similar to our research problem to provide good insight in terms of general framework, cue extraction, and inference methods. In the remaining of this thesis, we present our approach to analyze employment interviews and video resumes.

3 Collection of employment interviews

One of the objectives of this thesis is to investigate the formation of hirability impressions in employment interviews. To address this broad research problem, we collected the SONVB employment interview dataset, a corpus comprising 62 recordings of real employment interviews, where participants were applying for a paid marketing job. This data collection was motivated by the lack of available datasets including a job interview scenario. In this chapter, we present the experimental design and the technical setup used for this data collection. This dataset was collected within the framework of the SONVB (Sensing Organizational NonVerbal Behavior) project[1], a collaboration between Prof. Daniel Gatica-Perez (Idiap Research Institute), Prof. Marianne Schmid-Mast (University of Neuchâtel), Prof. Tanzeem Choudhury (Cornell University), and their respective research teams. The data collection was done in collaboration with Denise Frauendorfer (PhD student at University of Neuchâtel). A condensed version of this chapter was originally published in [97].

3.1 Scenario

Participants were recruited by advertising a part-time research assistant job opening in a social psychology lab. The position was a marketing job, where the hired applicants were expected to convince people on the street to participate to psychology experiments, and the job was paid 200CHF (208USD) for four hours of effective work. No specific requirement was set other than fluency in French, but applicants were expected to have strong communication, persuasion, conscientiousness, and stress resistance skills. The job offer was advertised on classified ads web platforms of two Swiss universities, and fliers were disseminated in three Swiss universities. Due to the large participation of students (90% Bachelor and Master students, 4.8% PhD students, 3.2% employed), the average age was 24 years (std = 5.68 years). The gender was somewhat unbalanced: 45 females (72.5%) and 17 males (27.5%).

A consent form was completed by participants upon arrival at the sensing lab, installed by Idiap Research Institute at the University of Neuchâtel. The informed consent form included

[1] www.idiap.ch/project/sonvb

In this corpus, we used a structured behavioral design, meaning that each interview followed the same structure and that some questions were related to applicant past experiences in specific situations. The sequence of questions is listed below:

1. Short self-presentation.
2. Motivation for applying to the job.
3. Importance of scientific research (which is the field of the job).
4. Past experience where communication skills were required.
5. Past experience where persuasion skills were required.
6. Past experience of conscientious/serious work.
7. Past experience where stress was correctly managed.
8. Strong/weak points about self.

Questions 4-7 are behavioral and were used to assess four hirability measures (communication, persuasion, conscience, stress resistance). Specifically, they were coded based on the quality of the applicant answers to these questions. One additional hirability measure (hiring decision) was coded on the whole interview sequence.

Figure 3.1 – Interview structure and hirability annotations.

the following points: (1) the interview was audio- and video-recorded; (2) data could only be used for research purposes; (3) data distribution outside the project team was not permitted for data privacy reasons; (4) videos, snapshots, or audio snippets could be used as demos or included in publications upon specific request; (5) participants could withdraw their consent at any time. Once signed, applicants were given a copy of the consent form. All applicants agreed to give their consent.

Applicants were then asked to complete a series of psychometric questionnaires. The questionnaires used in this data collection are presented in Section 3.3. The interview was designed as a structured behavioral interview, *structured* meaning that the interview strictly followed the same sequence of questions, ensuring that comparisons could be made between candidates, and *behavioral* implying that some questions were related to applicant past experiences in specific situations, eliciting a wide variety of behavioral responses. The psychology literature suggests that structured behavioral interviews are among the most valid tools for selecting applicants [69]. The interview structure is detailed in Figure 3.1. It included four behavioral questions, related to the specific skills required for the job, namely communication, persuasion, conscientiousness, and stress resistance, as well as standard job interview questions (self-presentation, motivation to apply for the job, strong/weak points about self). The interviews were dyadic: the interviewer and the applicant were seated at both ends of a table (see Figure 3.2). All interviews were conducted by the same person, the doctoral student in organizational psychology at University of Neuchâtel. The average interview duration was ~11 minutes; in total, the dataset comprises 670 minutes of recording.

3.2 Technical setup

Audio was recorded using a Microcone device [6], a commercial microphone array designed to record discussions of small groups. The Microcone records high-quality audio at a sample rate of 48kHz, and performs automatic speaker segmentation based on sound source localization. The device was placed on the table, at an equal distance between the applicant and the

Table 3.1 – Big-Five traits and related adjectives (taken from[59]).

Trait	Examples of Adjectives
Extraversion	Active, Assertive, Enthusiastic
Agreeableness	Appreciative, Forgiving, Generous
Conscientiousness	Efficient, Organized, Planful, Reliable
Neuroticism	Anxious, Self-pitying, Tense, Touchy
Openness to Experience	Artistic, Curious, Imaginative

interviewer (see Figure 3.2).

For video, two 1280 × 960 monocular cameras were used, recording both the interviewer and the job applicant synchronously at 26.6 frames per second, and frame timings were recorded at millisecond resolution. Camera views were quasi-frontal, filming the upper part of the body (see Figure 3.2). Audio-video synchronization was done manually by adjusting the delay between the pronounced words and the lip movements.

RGB-D data were recorded using two Kinect devices. Kinect is a device developed for the Xbox360 gaming platform, and has the ability to record depth in addition to standard RGB. Both the interviewer and the applicant were filmed in a quasi-frontal view. To record data from two Kinects, we developed a Python application based on the open-source libfreenect library [8] to interface the device and ffmpeg [3] to encode the frames. Kinect data were recorded at approximately 30 frames per second, and frame timings were recorded at millisecond resolution. As the Kinect recording application was developed during the data collection, only 43 interviews could be recorded with Kinects. The application was developed in collaboration with Kenneth Funes (PhD student at Idiap Research Institute).

3.3 Psychometric questionnaires

Job applicants were asked to fill in psychometric questionnaires before starting the interview. Three types of social constructs were assessed using questionnaires: personality, intelligence, and communication and persuasion skills.

3.3.1 Personality

For personality, we used the Big-Five model, which has received the most extensive support in psychology [59]. It represents personality at its highest level of abstraction and suggests that most individual differences in human personality can be classified into five empirically-derived bipolar factors, namely extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience (see Table 3.1). We used the NEO-FFI-R [43] questionnaire to assess the Big-Five personality traits. The questionnaire is standard and comprises 12 Likert items per factor ranging from 1 to 5 (half of the items are reversed), for a total of 60 items. The final score per dimension was obtained by averaging the items corresponding to each factor.

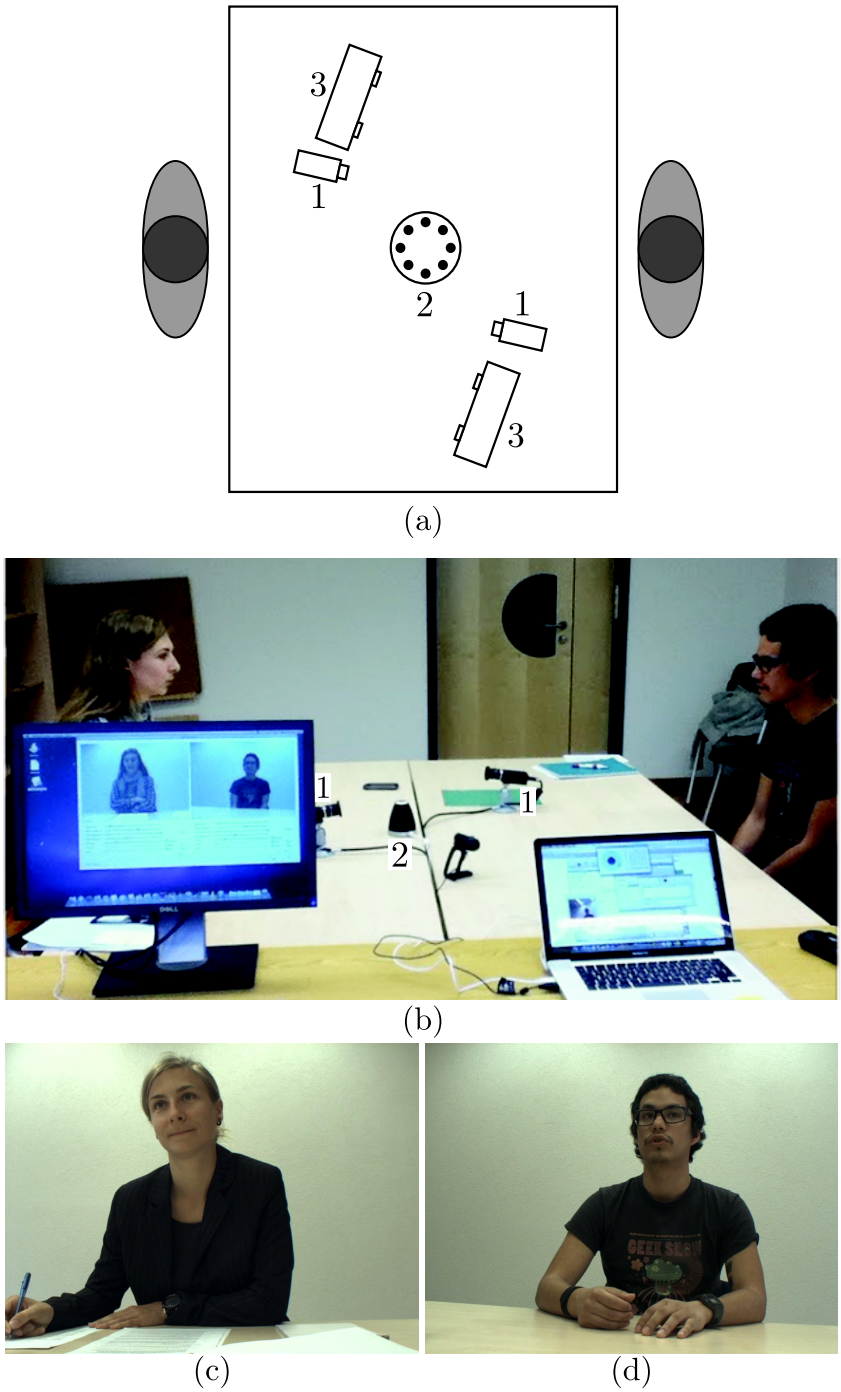


Figure 3.2 – (a) Sensor setup for the SONVB job interview data collection with (1) HD cameras, (2) Microcone microphone array, and (3) Kinect RGB-D devices; (b) Snapshot of the interview room; images of (c) the interviewer, and (d) the job applicant recorded by the HD cameras.

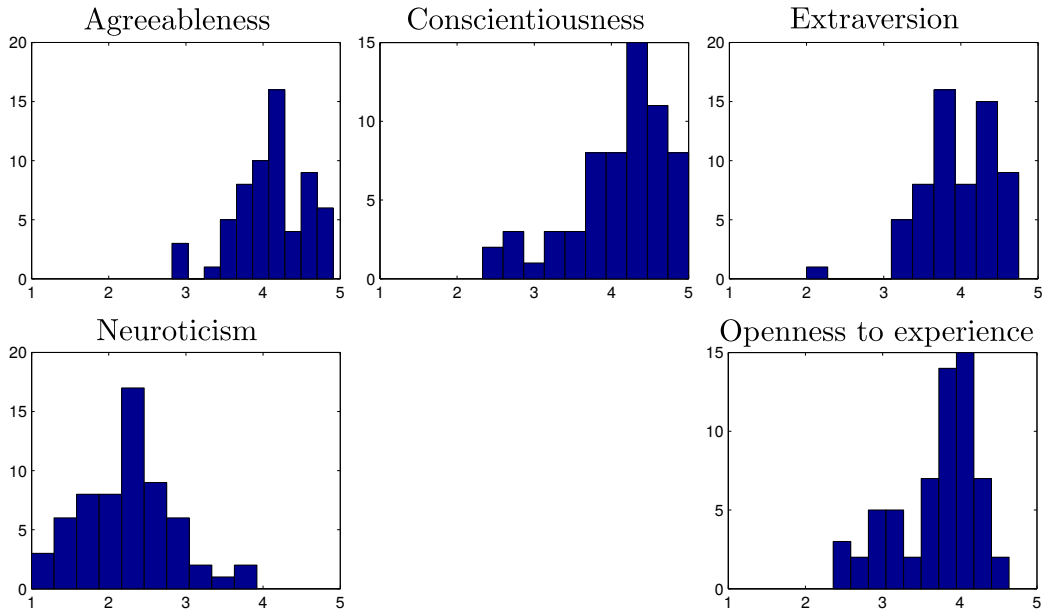


Figure 3.3 – Histograms for the Big-Five personality variables.

The histograms for the Big-Five personality variables for the SONVB corpus are displayed in Figure 3.3. We observe that most job applicants scored high on extraversion, agreeableness, conscientiousness, and openness to experience, and had low neuroticism scores. One hypothesis to explain this finding comes from the fact that personality traits were self-rated: applicants might have been positively biased by their representation of what an ideal candidate should be. Otherwise, all Big-Five variables roughly follow a Gaussian distribution, with a negative skewness for conscientiousness, extraversion, openness to experience, and agreeableness (respectively $-.98$, $-.97$, $-.69$, and $-.65$).

3.3.2 Intelligence

Intelligence is an important social construct in the personnel selection process. As one of the constructs consistently assessed by interviewers [69], it has been shown to correlate significantly with job performance across various types of job [69]. We used the Wonderlic Personnel Test [13] to assess applicant general intelligence. The test is standard and is composed of 50 questions to be answered within a time span of 12 minutes. The questions were designed to measure vocabulary, arithmetic reasoning, and spatial ability. The final intelligence score can be obtained by counting the number of correct answers.

3.3.3 Communication and persuasion

The job for which applicants were interviewed was a marketing job, which typically requires strong social skills such as communication and persuasion. To assess these skills, a questionnaire was designed by Prof. Marianne Schmid Mast's team (Université de Neuchâtel) based on

Table 3.2 – Pearson's pairwise correlation between hirability variables across annotation schemes ($^{*}p < .05$, $^{**}p < .01$, $^{\dagger}p < .001$). ($N = 62$).

Hirability variable	AV-FULL	AV-HR	FULL-HR
Communication	$0.90^{\dagger}$	0.35^{**}	0.31^{*}
Persuasion	$0.88^{\dagger}$	$0.48^{\dagger}$	0.41^{**}
Conscience	$0.84^{\dagger}$	0.33^{**}	$0.45^{\dagger}$
Stress Resistance	$0.89^{\dagger}$	$0.42^{\dagger}$	0.37^{**}
Hiring Decision	$0.94^{\dagger}$	$0.66^{\dagger}$	$0.66^{\dagger}$

Table 3.3 – Descriptive statistics of the HR hirability scores ($N = 62$).

Hirability variable	mean	std	skew	min	max
Communication	3.016	0.983	0.489	1	5
Persusasion	3.097	1.036	-0.105	1	5
Conscience	3.097	0.953	0.379	2	5
Stress Resistance	3.081	0.795	-0.144	1	5
Hiring Decision	6.161	1.803	-0.615	1	10

the Social Skills inventory [115]. Examples for items are: "In general I communicate in a clear manner" and "I often succeed in selling my point of view".

3.4 Hirability impressions

Hirability is a social construct which is dependent on the type of job, the content of the interview, and how the interview is conducted [89]. For this reason, there exists no standard way of assessing hirability. Moreover, there is no single definition of hirability, but it is rather composed of several scores related to the variables that interviewers and raters assess during the interview. In this study, five hirability scores were defined, four of which were specific to the four behavioral questions of the structured interview and the skills required for the job, while the remaining one was related to the full interview. More specifically, the abilities to communicate, persuade, work conscientiously, and resist stress (which were the qualities required for the job interview) were rated based on the quality of the applicant's response to the questions (see Figure 3.1 for more details). Additionally, the hiring decision score was annotated on the full interview. Each hirability score consisted of a score ranging from 1 to 5, except for the hiring decision which consisted of a score from 1 to 10.

Three rounds of annotations of hirability impressions were conducted:

1. **Audio-video hirability impressions (AV)**. For the first round of annotations, hirability impressions were annotated by a master's student in organizational psychology trained in recruiting applicants. The annotator was provided with the exact job description and job profile. She then watched the full job interviews and assigned the five hirability scores to each job applicant. For validity, a second coder (another trained organizational psychology master's student) rated 10 job interviews. Inter-rater agreement was good

Table 3.4 – Pearson's correlation between the HR hirability variables ($^{\star}p < .05$, $^{\star\star}p < .005$, $^{\dagger}p < .001$). ($N = 62$).

	1	2	3	4	5
1. Communication		$0.39^{\star\star}$	$0.26^{\star}$	$0.48^{\dagger}$	$0.59^{\dagger}$
2. Persusasion			$0.51^{\dagger}$	$0.43^{\dagger}$	$0.73^{\dagger}$
3. Conscience				$0.49^{\dagger}$	$0.63^{\dagger}$
4. StressRes					$0.70^{\dagger}$
5. HirDecision					

($r \in [.69, .99]$, using Pearson's correlation coefficient). This first round of annotations was conducted early on in the project, therefore was used as the basis for the computational analysis of hirability impressions (Sections 5.1, 5.2, 5.3, 5.5, and 5.6 of the dissertation).

2. **Full hirability impressions (FULL)**. To analyze the use of questionnaires for the prediction of hirability impressions (Section 5.4), it was necessary to assume that the recruiter also had access to the questionnaire data. To this end, we performed a second round of hirability annotations, where the coder started by looking at the questionnaire outputs before watching the full interview recordings. The annotations were completed by a master's student in organizational psychology trained in recruiting applicants. For inter-rater agreement, a secondary coder rated a subset of the data ($N = 10$) and the agreement was good ($r \in [.72, .93]$ using Pearson's correlation coefficient).

3. **Professional hirability impressions (HR)**. Later in the project, we had access to human resources (HR) professionals who were willing to give their impressions on how the applicants performed during the job interview. This third round of hirability impressions were annotated by a pool of five HR professionals, who watched the full video (and audio) of the job applicant. All interviews were annotated in total by three raters, who were provided with the job description and profile. Inter-rater agreement proved to be high among HR professionals, with $ICC(1,1) \in [.32, .52]$ and $ICC(1,k) \in [.59, .77]$, using the intraclass correlation coefficient. These annotations can be seen as the gold standard as the raters were professionally trained and had years of experience in hiring candidates. Later work of this thesis (Chapters 6) was based on HR hirability impressions.

Histograms for the HR hirability impressions are shown in Figure 3.4. Communication and stress resistance were observed to have strong modes. For the three other hirability variables, smoother distributions were observed. A notable observation is that the hiring decision variable spans from very low to very high. To avoid redundancy, histograms of hirability variables for the two other annotation schemes (AV and FULL) are not shown.

To understand the differences between the three types of hirability annotations, we computed Pearson's pair-wise correlations between annotation schemes for each hirability variable (Table 3.2). With correlation coefficients ranging from .84 and .94, almost no difference was observed between AV and FULL, suggesting that the questionnaire data did not affect much the formation of hirability impressions. Although significantly correlated, HR annotations

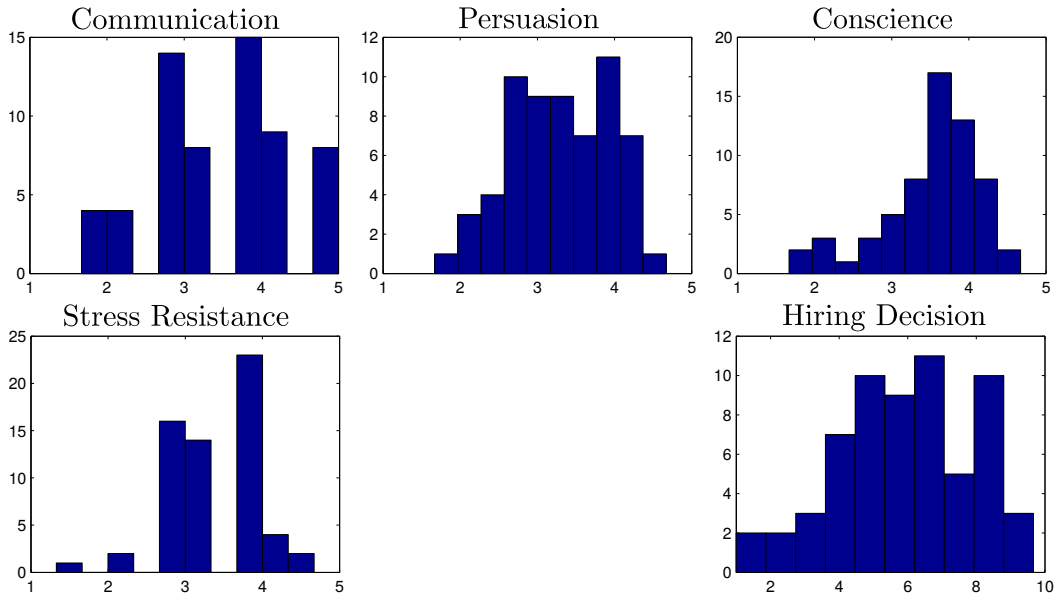


Figure 3.4 – Histograms for the HR hirability impressions.

showed lower correlations with the two other annotation schemes (AV-HR: $r \in [.33, .66]$; FULL-HR: $r \in [.31, .66]$). Interestingly, the hirability variable with the highest correlation was the hiring decision score. Assuming that the hirability scores given by the HR professionals are the gold-standard, these findings suggest that trained organizational psychology students were able to give hirability ratings similar to experienced HR professionals, but were not as accurate when it came to assessing more specific skills.

The descriptive statistics of the HR hirability scores are displayed in Table 3.3. The table shows a reasonable skewness for all the hirability scores. There was no need for transformation of the variables, because their distributions appeared to meet the assumptions of the statistical methods used throughout the thesis. The maximum possible value was reached for all hirability scores, while all except conscience reached the minimal possible value. The descriptive statistics for males and females were also computed, but no noticeable differences between gender were observed, therefore these results were not reported.

We present the pairwise correlations (using Pearson's correlation) between the HR hirability variables in Table 3.4. Note that the amount of shared variance between two variables can be obtained by taking the square of the corresponding correlation coefficient. We observe that all hirability scores were significantly correlated with each other. These correlation values suggest that the hirability scores used in this study were valid in the sense that they were measuring the same construct.

Table 3.5 – Summary of the SONVB data subsets.

Data subset	N	Hirability annotations	Sections
SONVB-NODS	8	-	4.1
SONVB-KIN	43	AV	5.5
SONVB-ALL	62		
SONVB-AV	62	AV	5.1, 5.2, 5.3, 5.6
SONVB-FULL	62	FULL	5.4
SONVB-HR	62	HR	6

3.5 Data subsets

Among the 62 job interviews collected for the SONVB dataset, 43 were recorded with the Kinect setup. Additionally, three different hirability annotation schemes were completed. Last, a test dataset of 8 natural dyadic interactions was collected prior to the recording of employment interviews. The data subsets used in this thesis are the following (see Table 3.5 for a summary):

1. **SONVB-NODS**. Consisting of 8 natural dyadic interactions, this data subset was collected before recording the job interviews as a test of the sensor setting. Dyads were acquainted and were instructed to have a discussion of their choice. This data subset was used for the development and evaluation of a multimodal method to detect head nods in natural interactions (Section 4.1). No hirability annotation was conducted on this data subset.

2. **SONVB-KIN**. This data subset comprises the 43 job interviews recorded with the Kinect devices. The hirability impressions were gathered using the AV scheme (annotations made by a master's student based on the audio-visual recordings of the employment interviews). This data subset was used in Section 5.5.

3. **SONVB-ALL**. This data subset comprises all 62 job interviews. Three rounds of annotations were completed:

    - **SONVB-AV**. Hirability impressions were gathered using the AV scheme (annotations made by a master's student based on the audio-visual recordings of the employment interviews). This data subset was used in Sections 5.1, 5.2, 5.3, and 5.6.
    - **SONVB-FULL**. Hirability impressions were gathered using the FULL scheme (annotations made by a master's student based on the questionnaire outputs and the audio-visual recordings of the employment interviews). This data subset was used in Section 5.4.
    - **SONVB-HR**. Hirability impressions were gathered using the HR scheme (annotations made by a pool of five HR professionals based on the audio-visual recordings of the employment interviews). This data subset was used in Chapter 6.

3.6 Conclusion

In this chapter, we presented the collection of the SONVB employment interview dataset, which served as a basis for the main part of this thesis. This data collection was motivated by the lack of a publicly available dataset including a job interview scenario. This dataset comprises 62 real job interviews, where participants were applying for a marketing job. Interviews were structured (the same sequence of questions were used), behavioral (some questions related to past experiences), and dyadic. Both the applicant and the interviewer were recorded using multiple modalities (audio, video, and Kinect). Psychometric questionnaires were completed by job applicants to assess their personality (using the Big-Five personality model), intelligence, and communication skills. Annotations of hirability impressions were collected using three different schemes, depending on the level of expertise of the raters (master's students in organizational psychology and human resource professionals) and the material available (audio-visual recordings and questionnaire data output). This dataset is the basis for the rest of the dissertation, except Chapter 7.

4 Extraction of behavioral cues

One of the objectives of this thesis is to investigate the use of state-of-the-art methods to automatically extract behavioral cues from job interviews. The goal of this step is to obtain an accurate representation (*i.e.*, a feature vector) of the dyadic interaction, which can be fed into a machine learning framework to infer high-level interview social variables. To get a full picture of the job interviews from a nonverbal standpoint, we extracted behavioral cues from both the applicant and the interviewer and from the audio and visual modalities; we then combined modalities and persons of interest to obtain multimodal and relational features. As rationale for selecting the behavioral features to be extracted, we searched the psychology literature for nonverbal cues that were shown to play a role in job interviews. We then used the available computational tools to extract the features of interest. We did not limit ourselves to the use of existing extraction methods: we developed a multimodal method to detect head nods based on the finding in psychology stating that head movements were conditioned by the speaking status of the person under analysis [62]. We showed that using the speaking status improved the detection accuracy of head nods. Although the great majority of behaviors were extracted automatically, we also completed manual annotations. This was especially the case for body posture and hand gestures, where the goal was to assess the feasibility of using such cues for the inference of key interview social variables in an ideal case. Finally, to complement nonverbal behavior we also extracted features related to verbal content. We used a state-of-the-art representation of verbal content, automatically extracted from manual speech transcriptions. Analyzing language style from accurate transcripts allows us to benchmark the use of verbal content for the inference of interview social variables.

In this chapter, we present our method to extract behavioral cues from dyadic employment interviews. Figure 4.1 presents a graphical summary of this chapter, where the data subsets used and the behavioral feature subsets extracted for each section are specified. In Section 4.1, we present a multimodal method to extract head nods in natural settings. In Section 4.2, we discuss the method used to extract interviewer and applicant nonverbal cues from both the audio and visual modalities. In Section 4.3, we present the extraction of additional visual cues, namely applicant body posture and gestures from manual annotations and a coarse

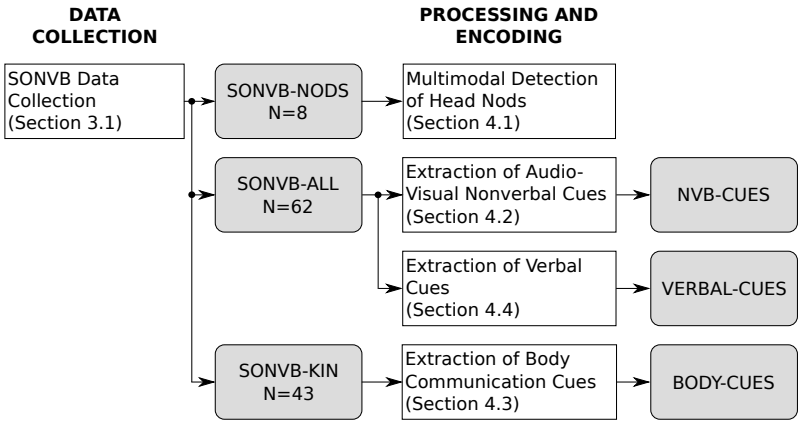


Figure 4.1 – Overview of Chapter 4.

representation of gestures. In Section 4.4, we present the method to extract behavioral features from verbal content, obtained from manual speech transcriptions. Finally, in Section 4.5 we summarize the extracted behavioral features and define the feature groups which were used in the next chapters. Section 4.1 was originally published in [99], Section 4.2 in [97], while Section 4.3 originally appeared in [98].

4.1 Multimodal detection of head nods

4.1.1 Introduction

In face-to-face interactions, head nods occur in every discussion. In most cases, people producing the head nods are not even aware of the social signal they emit: head nods are often the result of automatic processes. Independently of their function or meaning, head nods can be defined as vertical up-and-down movements of the head, rhythmically raised and lowered [62].

The social psychology community was the first to examine the functions of head nods during face-to-face interactions. Apart from the obvious function of signaling a 'yes', head nods are used *inter alia* to display interest, enhance communicative attention by occurring in synchrony with the other's speech, or anticipate an attempt to capture the floor (*i.e.*, signaling a turn claim) [62, 16]. Head nods form a major mode of communication in *back-channeling*, that is, during listener turns [62, 16]. Additionally, head nods can be used during speaker turns to elicit feedback from the listener [16]. The psychology literature suggests that the frequency of head nod events in face-to-face interactions can reveal personal characteristics or even predict outcomes. For instance, job applicants producing more head nods in employment interviews were reported to be often perceived as more employable than applicants who do not [57, 91]. In this sense, the ability to automatically detect head nods could be useful to build automatic inference methods of high-level social constructs.

Most methods for automatically detecting head nods have been developed in the context of human-computer interaction (HCI). The primary goal of these studies was to enable a machine to detect a 'yes' signaled by a head nod. Within this context, some studies proposed to track interest points of the face [75, 129] and use state-based approaches such as finite state machines [47] and hidden Markov models [96, 129] to detect nodding. These approaches show good performance in restricted contexts where the head motions are explicit. However, these methods do not allow to detect subtle head movements that occur quite often in natural human face-to-face interaction.

Nodding does not occur in a void. It is known that the speaking status of people influences the dynamics of the displayed head gestures [62]. When a person is speaking, the motion of his head has typically greater amplitude, larger frequency range, and follows a close to random pattern [62]. On the other hand, when the person is listening, his head tends to be more static as a result of both attention to the speaker and the fact of being silent. In this sense, head gestures are multimodal: the dynamics are conditioned on the speaking status of the actor.

The contextual nature of nodding has been used in the recent past. Work addressing the prediction of back-channel feedback makes use of these findings. The goal of this line of research is to enable robots or conversational agents to produce natural back-channels. In this type of setting, the contextual information such as lexical information or prosodic cues are used to predict head nods [95].

Furthermore, communicative contextual audio-based features have been used to improve the detection of head nods in dyadic scenarios [96, 94]. In [96], the scenario consisted of a human interacting with an embodied conversational agent; hence, head nods produced by the participant were not entirely natural. In [94], automatically and manually extracted contextual cues (prosodic and lexical features) related to the speaker were used to improve the detection of the listener's head nods. In this constrained dyadic scenario, only one person spoke while the other person was asked to listen silently.

While the value of using audio-based context from the perspective of the speaker to improve the detection of listener head nods has been established, one aspect that to our knowledge has not been studied in detail is the effect of the audio-based self-context on head nod detection. This section presents a multimodal method using the self-context to detect head nods in fully natural conversations where both protagonists freely interact.

4.1.2 Data and nodding annotations

In order to benchmark head nod detection methods, we used the **SONVB-NOD** data subset (see Section 3.5 and Figure 4.1), which consists of 8 natural interactions (16 videos treated individually). Pairs of participants were instructed to sit at both sides of a table and have a relaxed conversation on a topic of their choice. Dyads were acquainted before taking part to the experiment. In total, there were 9 different people (one person was present in all

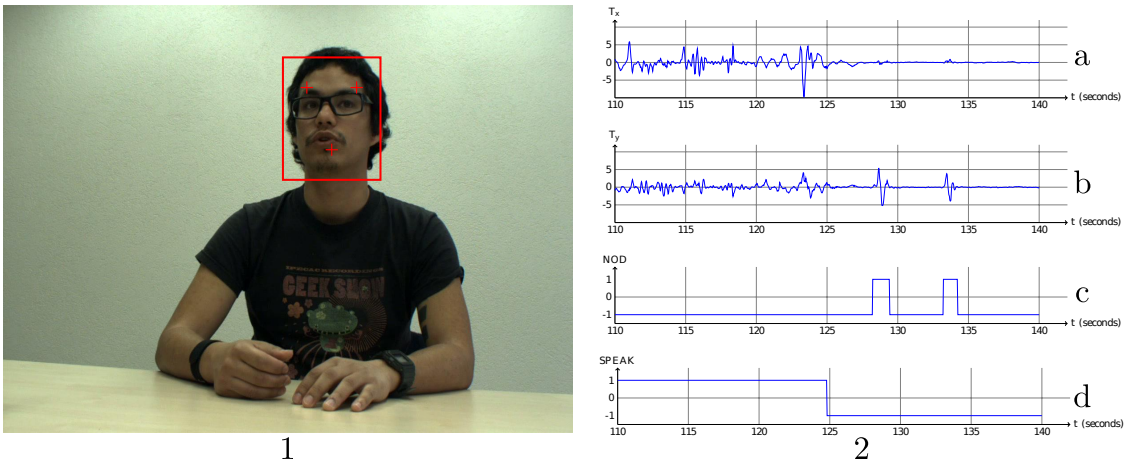


Figure 4.2 – Illustration of the motion estimation step. In (1), the red rectangle is the face bounding box provided by the face tracker; the three red crosses are the pre-defined points where the motion is computed using the parametric model of Equation 4.1. A 30-second sequence of the motion is displayed in (2); (a) and (b) are the estimated motion in the horizontal and vertical directions, respectively; (c) shows the sequence of annotated nods (1 = *nod*, −1 = *non-nod*); (d) shows the speaking status of the participant (1 = *speaking*, −1 = *silent*).

conversations).

In order to train and test the algorithm, annotations were performed on the dataset. Depending on the amplitude and duration of the up-and-down oscillatory movements, head nods can be difficult to code; two classes of head nods were therefore defined: *obvious* and *subtle*. Head nods were annotated by the PhD candidate, thus well acquainted with the concept of nodding, who noted the onset and offset time of an event, and qualitatively decided the nod class based on nod amplitude and duration. Speaking status was also manually annotated, marking the beginning and the end of a speaking segment. In total, nodding occurred during 858 seconds (22'812 frames), of which 92% occurred when the person under analysis was silent. Average head nod duration was 1.2 seconds.

4.1.3 Multimodal head nod detection

As stated in the introduction, head nods are defined as vertical up-and-down movements of the head rhythmically raised and lowered. This implies an oscillatory pattern in the vertical axis, while the motion in the horizontal axis is limited. In order to encode this effect, we constructed features based on fine-grain motion detection and transformation into the frequency domain. The extraction of the features to characterize head nods follows a similar spirit than [96], but relies directly on the motion estimates derived from the video sequence rather than on the output state of a head tracker, which might not be so sensitive to subtle movements of the head. A binary classifier is then used to assign frames to one of two classes, *nodding* and *not nodding*.

Motion Estimation

Given the bounding box output of a face tracker (using the method described in [114]), the goal is to detect the motion in the horizontal and vertical directions. To perform this task, we used a parametric motion model which estimates the best set of parameters between the previous and current frames, using the face bounding box region.

$$V(x_i, y_i) = \begin{bmatrix} v_x(x_i, y_i) \\ v_y(x_i, y_i) \end{bmatrix} = \begin{bmatrix} t_x + a_1 x_i + a_2 y_i \\ t_y + a_3 x_i + a_4 y_i \end{bmatrix}. \tag{4.1}$$

We used the affine motion model defined in Equation 4.1, where (x_i, y_i) denotes a point in the image, $V(x_i, y_i)$ is the flow vector modeled at point (x_i, y_i), and $t_x, t_y, a_{1:4}$ are the model parameters. Visual motion estimation over the whole face region provides an accurate estimation of head movements and therefore allows to capture subtle patterns using a multiresolution robust estimation method [102]. Parameters t_x, t_y, $a_{1:4}$ were estimated using least-mean-squares, implemented by the software package Motion2D[1].

We then computed the velocity at three arbitrarily defined points (see Figure 4.2) inside the bounding box, using the parameters of the optical flow model (Equation 4.1), providing the horizontal and vertical components of the motion at these three points. Roughly speaking, these points are around the mouth and eyes of the participant. The use of three points is equivalent to the affine motion model (6 parameters), and captures rotational and lateral movements of the head, which would not be the case of a single point; moreover, using more points would not provide any additional information. Compared to directly using the model parameters, using the motion estimated at specific points shows the advantage of not requiring any type of weighting or transformation as all values refer to the same physical quantities (*i.e.*, 2-D displacement) and share the same units (pixels).

Typical motion time-series for speaking, nodding while silent, and not nodding while silent are illustrated in Figure 4.2. The figure illustrates that head nod activity does present differences depending on the self-speaking status, and that building nodding models separately for the speaking and silent cases could reduce the confusion between, for instance, a head nod and a quasi-random head gesture displayed during a speaking turn.

Frequency Domain Analysis

Given the head motion in the horizontal and vertical directions, the goal is to capture the oscillatory characteristics of a head nod. In order to perform this task, we applied a Fourier transform (with Gaussian temporal window) to the velocity vectors $(v_x(x_i, y_i), v_y(x_i, y_i))^T$ of the pre-defined points, considering each vector component as an independent time-series.

[1] `http://www.irisa.fr/vista/Motion2D/`

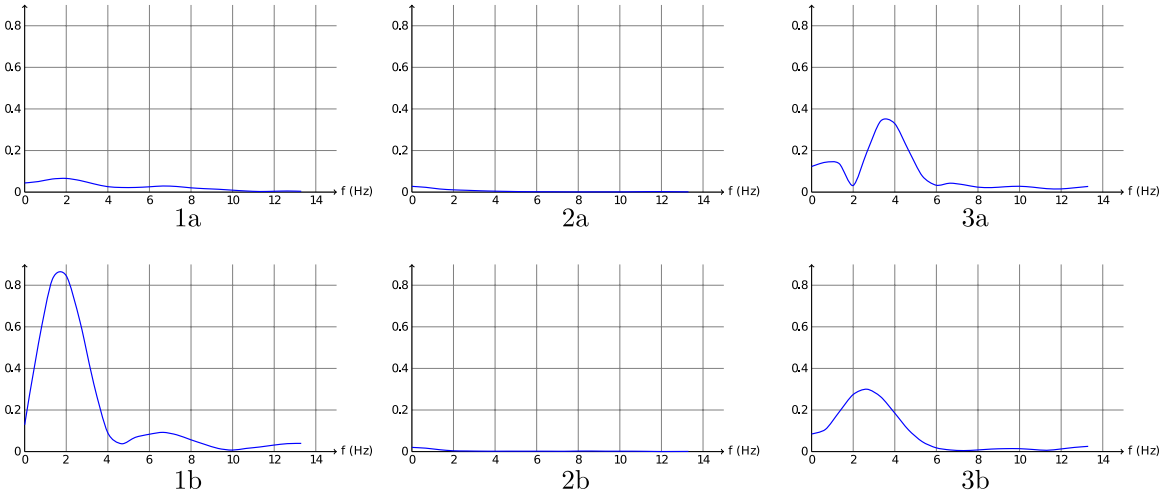


Figure 4.3 – Examples of typical Fourier transform outputs. (a) and (b) are the Fourier outputs taken on the temporal motion sequence in the horizontal and vertical direction, respectively. Referring to Figure 4.2, (1) is a Fourier sample taken at $t = 128.8s$ (*nod, non-speaking*); (2) is sampled at $t = 131.4s$ (*non-nod, silent*); (3) is taken at $t = 116.2s$ (*non-nod, speaking*).

Although this independence assumption does not strictly hold, the computation is greatly simplified. Finally, the feature vectors are constructed by concatenating the Fourier transform outputs. Typical Fourier features for speaking, nodding while silent, and not nodding while silent are illustrated in Figure 4.3.

Classification

The goal is to assign each feature vector to either of the two classes, *nodding* or *not nodding*. To this end, we used a linear support vector machine (SVM) as defined in Equation 4.2, where y denotes the output class label, $\mathbf{x}$ indicates the feature vector, $\mathbf{w}$ represents the SVM weights, and b is the SVM bias. For each of the speaking status values, we trained a separate linear SVM to perform the classification. This multimodal approach directly takes into account the switching dynamics of head movements, depending on the speaking status of the person under observation, as suggested by previous work in psychology [63].

$$y = sign(\mathbf{b}^T\mathbf{x} - b). \tag{4.2}$$

The training set was defined as follows. The positive set was composed of all frames labeled as *obvious nods*. The negative set was selected randomly from the set of frames labeled as *non-nod*. Frames labeled as *subtle nods* were not used for training because they can be too similar to *not nodding* features. In addition to this, transitional frames were discarded to attenuate the dependence to time-related annotation inaccuracy. The training set was balanced, *i.e.* the

number of positive and negative examples was equal. Approximately 5000 training examples were used for each class. We further segmented the data into *speaking* and *silent*, training each separate model on its own training set.

To validate our hypothesis that self-context in terms of speaking status improves nodding detection, we implemented a baseline method using the visual modality only. For the visual-only method, we used a single SVM trained on the full training set (*i.e.* not separating it into *speaking* and *silent*).

4.1.4 Results

The evaluation of the head nod detection method was conducted at the frame level. Leave-one-out cross validation was performed at the sequence level: the algorithm was trained on all except one sequence and tested on the remaining one. The binary output of the SVM classifier was compared to the annotated ground truth (including *obvious* and *subtle* nods).

In Figure 4.4, we display the receiver-operating characteristic (ROC) curve for both the visual-only and the multimodal approaches. The ROC is obtained by plotting the true positive rate against the false positive rate at various threshold settings, and illustrates the performance of a binary classifier. The F_1 score, defined by the harmonic mean of precision and recall (Equation 4.3), was also computed.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}. \tag{4.3}$$

The multimodal method significantly outperforms the visual-only one: $F_1^{visual} = 0.559$, $F_1^{multi} = 0.628$. These results show that using the self-speaking status of a person improves the detection rate, highlighting the difference in dynamics of head gestures between *speaking* and *silent* suggested in the psychology literature [62]. This result confirms previous findings [96, 94] that have shown the advantage of using audio-based contextual cues for this task, with the novel angle that using self-context (as opposed to interaction partner-based context) is also advantageous. Moreover, independently of the approach (multimodal or visual-only), the method we developed to extract head nods in natural settings yielded competitive results on this dataset.

Head nods of small amplitude and duration (*subtle* nods) were in general accurately detected by the proposed method. Additionally, because of the switching dynamics conditioned on the speaking status, the number of false positives was kept low. In Figure 4.5 we display the learned SVM parameters for both the *silent* and *speaking* classifiers. For the *silent* case, we first observe that nods were characterized by a vertical oscillatory pattern with a frequency ranging roughly from 1 to 6 Hz. The SVM weights corresponding to the horizontal motion were close to zero, weakly affecting the classifier output. For the *speaking* case, only weights corresponding

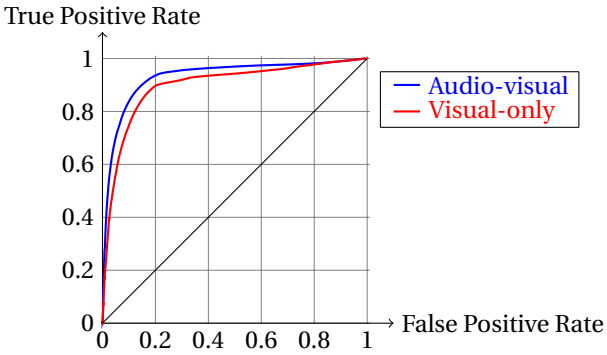


Figure 4.4 – Receiver-operating characteristic curve (ROC) for the two head nod detection approaches. In red: visual-only method. In blue: audio-visual approach.

to vertical oscillatory motions ranging from 3 to 5 Hz were positive, but with lower values than for the speaking case. Furthermore, the bias for the *speaking* classifier ($b_{speaking} = 0.609$) was much greater than for the *silent* one ($b_{silent} = -0.019$), further attenuating the detection of head nods when the person was speaking. In other words, *speaking* can be seen as an attenuation factor of the head nod detector.

4.1.5 Conclusion

In this section, we developed and evaluated a multimodal method to detect natural head nods in face-to-face interactions. Our work examined in detail the effect of the speaking self-context on head nod detection. Two nodding models were trained, depending on the speaking status of the person under analysis. Compared to the baseline vision-only method, the results demonstrated that audio-based self-context improved the detection of head nods, underlining the difference of head gesture dynamics conditioned on the speaking status of the person, as suggested by previous work in psychology. The developed method yielded competitive results on this dataset, allowing to detect subtle nods while keeping the number of false positive low.

4.2 Audio-visual nonverbal cues

In this section, we present the nonverbal behavioral cues extracted from the SONVB interview dataset (**SONVB-ALL**, see Section 3.5 and Figure 4.1). We extracted nonverbal features from the audio and visual modalities. We built multimodal and relational features by combining unimodal features. As a rationale for selecting the behavioral features to be extracted, we searched the psychology literature for nonverbal cues which were shown to play a role in job interviews. We then used available computational tools to extract the features of interest. As the interviewer's nonverbal behavior has been shown to have an impact on the interview outcome [45], we extracted behavioral cues from both the applicant and the interviewer. The feature subset corresponding to audio-visual nonverbal cues is referred to as **NVB-CUES**.

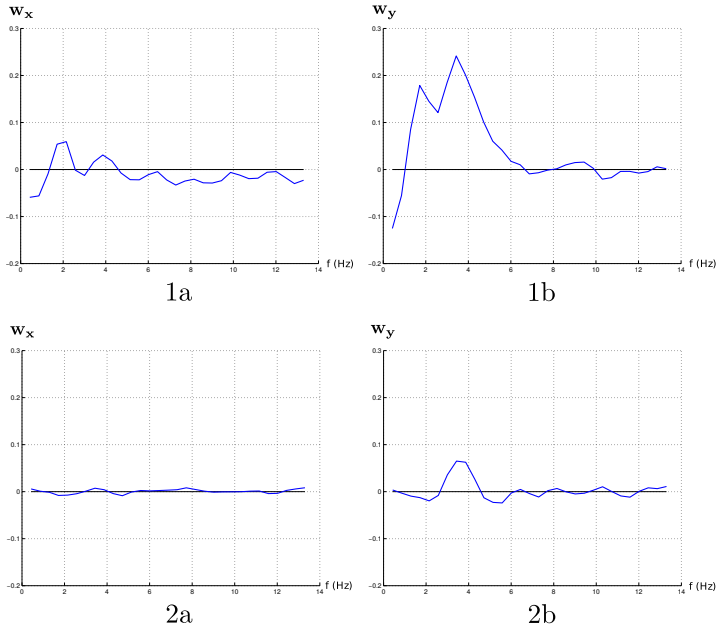


Figure 4.5 – Learned SVM weights for the *silent* (1) and *speaking* (2) classifiers along the horizontal (a) and vertical (b) axes, $b_{silent} = -0.019$, $b_{speaking} = 0.609$.

4.2.1 Audio cues

Speaking activity

Cues based on speaking activity such as applicant pauses [45], speaking time [57], and speech fluency [90, 45] have shown effects on interview ratings. All speaking activity cues were based on the speaker segmentations given by the Microcone [6]. The device, in addition to recording the audio at 48 kHz, has the ability to automatically segment speaker turns, using a filter-sum beamformer followed by a post-filtering stage. The resulting speaker segmentations were stored in a file containing the relative time (start and end) and the speaker identifier. The objective performance of the speaker segmentation was not evaluated, but upon manual inspection we found that the number of errors was low. The following speaking-activity-based features were extracted for the interviewer and the applicant:

- **Speaking time**. Total speaking time was extracted by adding all speaking turn durations. The number was then normalized with respect to the average interview duration.
- **Speaking turns**. Speaking turns were defined as speaking segments longer than 2 seconds. Speaking turns were merged if the non-speaking gap between them was shorter than 2 seconds. The number of speaking turns, average turn duration, turn duration standard deviation, and maximum turn duration were used as behavioral features from the speaking turns.
- **Pauses**. The aforementioned non-speaking gaps shorter than 2 seconds were defined as pauses. The number of pauses was normalized with respect to the average interview duration.

- **Short utterances**. Short utterances were defined as speaking segments of duration smaller than 2 seconds. The number of short utterances was normalized with respect to the average interview duration.

Prosody

Applicant prosodic cues (*i.e.* pitch, speaking rate, and energy) were found to be significantly correlated with job interview outcomes in several psychology studies [90] [45]. From the speaker segmentations, we obtained the speech signals for the interviewer and the applicant, from which we extracted the energy, the perceived fundamental frequency, and the voiced rate (number of voiced segments per second). Methods for extracting prosodic cues are well documented (*e.g.*, [24]), and we used the speech feature extraction code from the Human Dynamics Group at the MIT Media Lab [7]. For speech energy, pitch, and voicing rate, we extracted the following statistics: mean, standard deviation, minimum, maximum, entropy, median, and quartiles.

4.2.2 Visual cues

Organizational psychology literature suggests that visual cues are often used by interviewers to assess the applicant's hirability in job interviews. Gaze, smiles, hand gestures, head gestures, posture, and physical attractiveness were found to have a significant effect on hirability ratings [20, 45, 57]. We decided to automatically extract a smaller number of cues including head nods, overall visual motion, and face-region optical flow. In addition to these three visual cues, we manually coded applicant gaze and smiles. In the following, we present the method used to extract these visual cues.

Head nods

Head nods are defined as vertical up-and-down movements of the head, rhythmically raised and lowered. We used the method proposed in Section 4.1 [99] to automatically extract head nods. From the detected nods, we recorded the number of nods and total nodding time. These numbers were then normalized with respect to the average interview duration.

Head region visual motion

This cue quantifies the amount of head motion displayed by a person, and was based on the parametric optical flow estimation method described in [102]. The overall optical flow between two consecutive frames was computed inside the face bounding box, using a parametric affine model (Equation 4.1). The estimated model was then used to compute the motion at three predefined points within the bounding box, as discussed in Section 4.1.3 (see Figure 4.2). We then took the average motion of these three points, and extracted the absolute value of the

Figure 4.6 – Example of applicant weighted motion energy image (WMEI), quantifying the overall visual motion during an interview. Each pixel intensity indicates the visual activity at its position.

horizontal and vertical velocity components, and computed the velocity magnitude. The mean and standard deviation of these values were used as features.

Overall visual motion

This feature quantifies the amount of visual movement displayed by the applicant and interviewer during the job interview and is an indication of kinetic expressiveness. We used a modified version of motion energy images, called Weighted Motion Energy Images (WMEI) [27] which summarizes the motion throughout a video as a single grayscale image, where each pixel intensity indicates the visual activity at its position. An example of WMEI is shown in Figure 4.6. From the WMEIs, we computed statistical features as descriptors of overall visual motion: mean, median, standard deviation, minimum, maximum, entropy, quartiles, and center of gravity.

Smiling

This cue was manually annotated by a social psychologist who counted the number of smiling events for the applicant. This number was then normalized with respect to the average interview duration. A second annotator coded smiles in a subset of the dataset ($N = 10$), and interrater agreement was high ($r = .95$).

Gazing

An organizational psychologist manually coded the percentage of time for which the candidate was looking at the interviewer. A secondary annotator coded the same value on a subset of the data ($N = 10$), and interrater agreement was high ($r = .95$).

Physical appearance

To assess the applicants' attractiveness, three variables were coded by 10 raters based on still images: physical attractiveness, sympathy, and appreciation. Annotators were asked to answer the following questions: "How attractive do you find this person?" for physical attractiveness, "How sympathetic do you find this person?" for sympathy, and "How much do you appreciate this person in general?" for appreciation. Each rater gave a grade between 1 (low appreciation) and 5 (high appreciation), and the average over all raters was taken for the three variables.

4.2.3 Audio-visual and relational cues

Social computing studies have demonstrated the predictive validity of multimodal and relational features. For instance, features such as "looking-while-speaking" [29], cues related to the group [73] or to the dyad [44] have been successfully used for the automatic inference of social constructs. To encode the multimodal and relational characteristics of nonverbal behaviors, we combined audio and visual cues, as well as cues related to the applicant and to the interviewer. The rationale for combining two binary sequences is illustrated in Figure 4.7, and comprised two steps. First, the binary time-series were dilated using parameter τ, in order to account for the slight asynchrony between two co-occurring audio-visual or relational events. Second, the two dilated binary time-series were combined by applying a logical AND operator to each frame of the time-series. The following multimodal/relational behavioral features were extracted:

- **Audio back-channeling**: events when a person produced a short utterance while the other was speaking.
- **Visual back-channeling**: events when a person nodded while the other was speaking.
- **Audio-visual back-channeling**: events when a person nodded and produced a short utterance, using dilating parameter $\tau \in \{0, 0.5, 1, 1.5, 2\}$ seconds to account for slight asynchrony, while the other was speaking.
- **Nodding while speaking**: events when a person nodded while speaking.
- **Mutual short utterances**: co-occurring events when the two protagonists produced a short utterance, using dilating parameter $\tau \in \{0, 0.5, 1, 1.5, 2\}$ seconds to account for asynchrony.
- **Mutual nods**: co-occurring events when the two protagonists nodded, using dilating parameter $\tau \in \{0, 0.5, 1, 1.5, 2\}$ seconds to account for asynchrony.

For each of these definitions, the total time and the number of events were stored as features. The numbers were normalized with respect to the average interview duration.

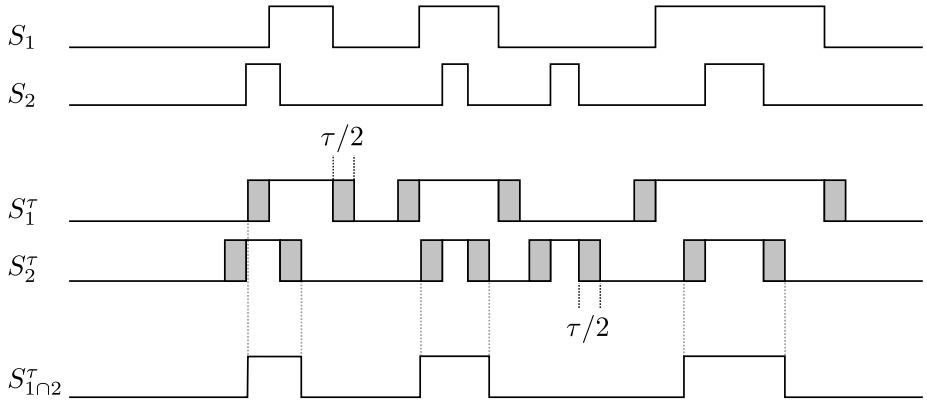


Figure 4.7 – Illustration of the scheme used for combining two time-series. S_1 and S_2 are the original time-series; S_1^τ and S_2^τ are the τ-dilated time-series. $S_{1\cap 2}^\tau \triangleq S_1^\tau \cap S_2^\tau$ is the resulting time-series.

4.3 Body communication cues

Body communication comprises what the face, head, eyes, limbs, and trunk transmit. Although the importance of head gestures, facial expressions, and gaze has been demonstrated in the literature, we focus here on the analysis of body posture and gestures. Gestures are an essential component of body communication as they are used to enrich the vocal content and aid listener comprehension by augmenting the attention, activating images or representations in the listener's mind, and increasing the recall of what is being said [78]. Moreover, restraining people from gesturing strongly affects the speakers' fluency [78]. Body posture is another important component of body communication; various emotions such as fear, sadness, or happiness have been shown to be correctly inferred from a person's pose [78]. In conversations, body postures can be used as markers during a conversation: for instance, changes of body posture can precede a long utterance and may be kept for the duration of the speaking turn [78]. Both gestures and postures are inherently multimodal, in that they do not only occur in the visual modality, but are conditioned on the speaking status (*i.e.*, audio modality) of the person. For this reason, we believe that it is necessary to consider the speaking status when analyzing posture and gestures.

Because posture and gestures are difficult nonverbal features to extract automatically, we started by extracting body communicative cues based on manual annotations, with the goal of benchmarking the use of body cues for the inference of organizational constructs in an ideal case, *i.e.* assuming that a perfect body posture detector existed. Then, we computed automatic visual features as a raw representation of hand gestures to compare with the ideal case. Furthermore, the speaking status provided by the Microcone was included in the extraction process. This section describes the method used to obtain multimodal body communication cues. Body cues were extracted from the **SONVB-KIN** data subset (see Section 3.5 and Figure 4.1) for the applicant only. The feature subset corresponding to body communication cues is referred to as **BODY-CUES**. The work in this section was done in collaboration with Alvaro

HH HT GT G ST

Figure 4.8 – Class examples. From left to right: hidden hands (HH), hands on table (HT), gestures on table (GT), gestures (G), self touch (ST).

Marcos Ramiro (PhD student at University of Alcala, Spain).

4.3.1 Annotations of body activity

Five classes were defined based on their occurrence in the dataset and their relevance in the nonverbal communication literature [78]: hidden hands, hands on table, gestures on table, gestures, and self-touch. They constitute an approximation for the applicant's body posture and gestures. Applicants were seated, therefore the posture was in a large part defined by the position of their arms. Other posture classes such as leaning forward or backward were also considered, but were discarded as the observed variability of such postures was low. Class definitions and examples can be seen in Table 4.1 and Figure 4.8, respectively.

Applicant body activity was annotated at the frame level by one person, with the help of a purpose-built script. To reduce the amount of frames to label, annotations were made every 15 frames (0.5 seconds); this temporal resolution was sufficient as no missing labels were observed while playing the full video at regular speed. To further reduce the amount of frames to label, we applied a motion threshold to the videos and annotated frames only when sufficient movement was present; unannotated frames in between annotated ones were assigned the same label as the latest annotated frame. This procedure allowed us to reduce the number of frames to annotate by 35%. In total, over 23000 frames were labeled. In order to assess the reliability of the annotations, a second person annotated 63 minutes of the dataset (≈5000 frames), and interrater agreement was satisfactory (Cohen's Kappa: $\kappa = 0.81$).

The class distribution of the corpus is shown in Figure 4.9. We observe that *hands on table* accounted for more than half of the labels. The dataset was recorded in a real setting, therefore it reflects the natural tendency of the participants while being seated. It should be noted that in 34.2% of the data the subject was silent while listening to the interviewer. Our proxy for beat gestures (*gestures* and *gestures on table*) were present 33.6% of the time. The least represented class was *hidden hands*, while *self-touch* appeared almost as often as *gestures*.

Table 4.1 – Class descriptions.

Class	Description
Hidden hands (HH)	No hands visible in the image
Hands on table (HT)	Resting the hands in the table
Gestures on table (GT)	Gesturing while the hands are close to the table, or the arms resting on it
Gestures (G)	Gesturing while the hands are not close to the table
Self-touch (ST)	Touching face, hair, or torso with one or both hands

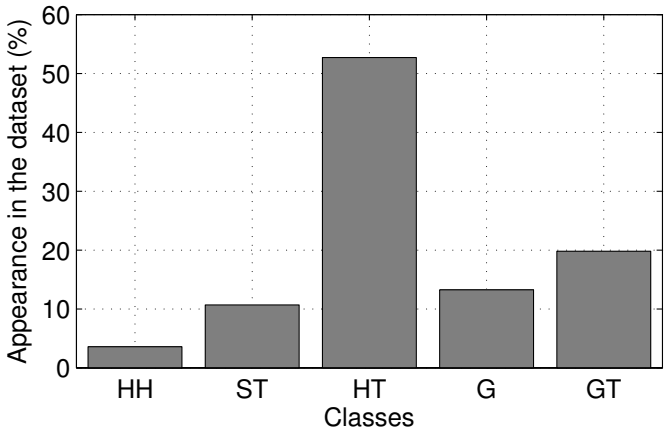


Figure 4.9 – Frequency of each class in the dataset.

4.3.2 Automatic features

Hand speed

To obtain estimates of hand speed, we used the method presented in [88] to compute the hand likelihood map for each frame of a video. This method makes the assumptions that the hands are the fastest moving parts of the video, that they are not the face, and that they have skin color. Based on these assumptions, the hand likelihood map can be computed as the product of a dense optical flow map, a binary face-mask image, and a skin-color segmentation binary image, all of which are automatically computed according to algorithms described in details in [88]. We implemented a simple but effective method to obtain the hand speed image: we multiplied the hand likelihood map with the pixel frame-difference and normalized it by the distance between the head and the table to account for variations in the camera placement. An illustration of the procedure to compute the hand speed image is displayed in Figure 4.10. As a last step, we obtained the hand speed energy $e_H(t)$ by summing all pixels of the hand speed image, resulting in a single value for the hand speed estimate for each frame of a video.

Image activity histograms

In order to obtain information about the hand position of the participants, we created an image activity descriptor along the vertical axis of the image. We defined a 12-bin histogram $\mathbf{h}_{OF}(t)$, which accumulates energy in different height bands of the dense optical flow image

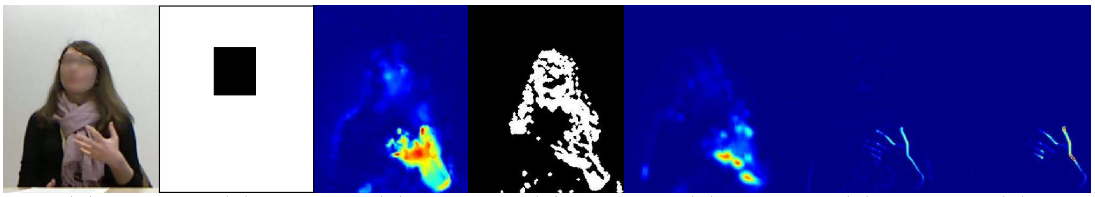


Figure 4.10 – Illustration of the hand speed image computation: (1) original image, (2) face mask, (3) optical flow map, (4) skin-color segmentation image, (5) hand likelihood map, (6) frame difference, and (7) resulting hand speed image.

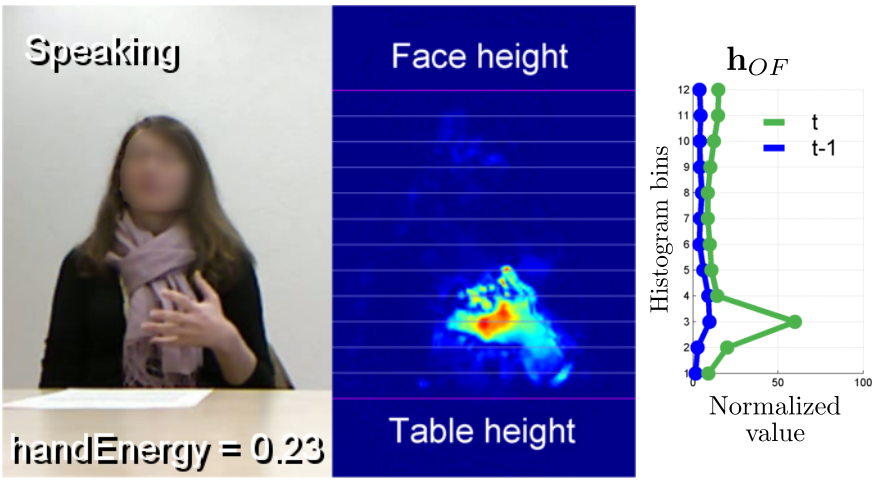


Figure 4.11 – Left: input image with overlaid speaking status and hand energy value. Center: Dense optical flow and image height division. Right: Activity histograms $\mathbf{h}_{OF}$.

(normalized by the distance between the table and the head). The histogram is able to capture two important factors which condition the applicant's visual activity, *i.e.* hand speed and hand height, which makes this feature suitable for the analysis of seated participants. Moreover, as the method is based on dense optical flow, it is appearance invariant, which makes it suitable for the analysis of subjects with different skin colors. An illustration of the image activity histogram can be seen in Figure 4.11.

4.3.3 Nonverbal cue encoding

We now describe the method to encode the nonverbal cues from the manual annotations of body activity, the speaker segmentations, the hand speed estimates, and the image activity histograms.

Figure 4.12 – Illustration for cues based on annotations of body activity. There are two HH events, six GT events, nine HT events, and no ST events. Statistics are computed from event durations. If no event occurred, the statistics are set to zero.

Cues based on annotations of body activity

Nonverbal cues were extracted from the manual annotations of body activity. To capture a "big picture" of the body activity, they were based on statistics derived from event durations. Events were defined as a sequence of frames in which the applicant showed the same type of body activity, and were characterized by their starting time and duration (see Figure 4.12). For all the activity classes, we computed the number of events, mean, median, standard deviation, lower and upper quartiles, minimum, maximum, range, position (in time) of shortest and longest events, and total relative time. It should be noted that it was possible for a given class to be missing in a given sequence. We addressed this by introducing a binary variable indicating whether at least an event occurred or not. The statistics on turn durations were set to zero if no event occurred. The list of body communication cues based on manual annotations is included in Table 4.2.

Table 4.2 – List of manual and automatic body nonverbal cues. Each cue was computed for the unimodal case (*i.e.* not taking the speaking status into account), and also for the speaking, silent, and aggregated cases (*i.e.* aggregating unimodal, speaking, and silent).

Manual features		
Posture class	**Statistics**	**Speaking status**
Hidden hands (HH) Self-touch (ST) Hands on table (HT) Gestures on table (GT) Gestures (G)	exists, mean, median, std, quartiles, # of events, min., max., range, rel. time, pos. of min./max.	Unimodal, Speaking, Silent, Aggregated

Automatic features		
Time-series	**Statistics**	**Speaking status**
Hand velocity (HV) Hand acceleration (HA)	mean, median, std, quartiles, zero-crossing rate, min., max., range, prop. non-zero	Unimodal, Speaking, Silent, Aggregated
Histograms (mode)	**Statistics**	**Speaking status**
Hand velocity histogram (HVH) Hand acceleration histogram (HAH)	mean, median, std, quartiles, zero-crossing rate	Unimodal, Speaking, Silent, Aggregated

Cues based on hand speed and activity histograms

The hand speed approximation $e_H(t)$ and the image activity histograms $\mathbf{h}_{OF}(t)$ (Section 4.3.2) not only provide information on *how much* hand movement occurred at a given instant, but also on *where* these hand movements occurred. We also extracted nonverbal cues based on those activity descriptors. To account for short bursts of hand movement characterized by quick changes of hand speed (which could be associated with beat gestures), we computed the hand acceleration. We defined the global hand acceleration at time t as $a_H(t) = |e_H(t) - e_H(t-1)|$, and the height-dependent acceleration as $\mathbf{a}_{OF}(t) = |\mathbf{h}_{OF}(t) - \mathbf{h}_{OF}(t-1)|$.

To extract nonverbal cues from the univariate time series e_H and a_H, we computed the mean, median, standard deviation, minimum, maximum, range, quartiles, proportion of non-zero elements, and zero-crossing rate. We also computed statistics related to the main mode of the histogram (*i.e.*, the position of the maximum histogram bin) to account for hand position. This includes the mean, median, standard deviation, quartiles, and zero-crossing rate. Table 4.2 shows the list of the automatically extracted body communication cues.

Exploiting the speaking status

To exploit the finding in psychology stating that body communication is conditioned on the speaking status [92, 78], we computed the statistics on manual body activity event duration, hand speed and acceleration, and activity and acceleration histogram modes for four different cases: (1) the unimodal case, *i.e.*, without taking into account the speaking status, (2) the speaking case, *i.e.* only using frames for which the applicant was speaking, (3) the silent case, and (4) the aggregated case, *i.e.* aggregating the three previous cases. Table 4.2 shows the list of all the body communication cues extracted.

4.4 Verbal cues

The words we use while speaking or writing reveal many aspects of our identity [105, 38]. Although previous work in psychology has shown that language style was associated with various personal social constructs, such as personality, affective states, status, deception, demographics (gender, age), or culture [38], little is known about the role of verbal content in the formation of hirability impressions. Among the few studies investigating the role of language style in the employee selection process, nonverbal behavior conditioned by the level of verbal content predicted high ratings in simulated job interviews [112].

Recent studies in social computing have investigated the use of verbal content for the prediction of social constructs in natural interactions. For instance, emergent leaders in small groups [121], or personality impressions of YouTube video bloggers [31] were predicted to some degree from manually annotated transcriptions. To the best of our knowledge, no computational study has investigated the role of language use on hirability in job interviews.

To examine the role of verbal behavior in job interviews, we extracted behavioral features based on verbal content. Specifically, we aim at assessing the use of linguistic and paralinguistic categories obtained from manual transcripts for the prediction of hirability impressions in job interviews. In this section, we present the verbal behavioral cues extracted for the job applicant from the SONVB interview dataset (**SONVB-FULL**). The feature subset corresponding to LIWC categories is referred to as **VERBAL-CUES**.

4.4.1 Manual transcriptions

Each job interview was transcribed by a master's student in organizational psychology, who was also a native French speaker. Each answer by an applicant was transcribed. Because the interviewer's questions did not vary across interviews, they were not transcribed. Below is an example of transcription for the first interview question (self presentation):

Original transcript (French)	**Translation**
Alors je m'appelle P. G., j'ai vingt-quatre ans, je fais des études en master à l'Université de ... en français en linguistique. Personnellement ce que je peux dire sur moi c'est que je suis une personne très communicative, que j'ai pas mal c'est vrai que d'amis assez large, assez diversifié. Et que ce qui m'intéresse principalement c'est les êtres humains, donc toutes leurs facettes.	*So my name is P. G., I am twenty-four, I am a Master's student at the University of ... in French in linguistics. Personally what I can say about myself is that I am a very communicative person, that I have, that's right, quite a few friends, quite broad, quite diversified. And that what mainly interests me is human beings, therefore all their facets.*

The average transcribed number of words was 931 for the full interview, ranging from 285 to 1716 words.

4.4.2 Linguistic Inquiry Word Count (LIWC)

As a high-level representation of verbal content, Linguistic Inquiry Word Count (LIWC) is a software module resulting from years of social psychology research, focused on the validation of the psychometric properties of a word categorization system that relates linguistic and paralinguistic categories to psychological constructs [130]. In its original English version, LIWC is built based on a dictionary of 4500 words and word stems. Each word found in the dictionary is assigned to one or several categories, belonging to five meta-categories, namely linguistic processes, psychological processes, personal concerns, spoken categories, and punctuation. Examples of categories for each meta-category are displayed in Table 4.3. Each word from the transcripts found in the dictionary is assigned to one or more categories, and increments the count of the respective category.

To obtain a representation of verbal content, we used the French version of the Linguistic Inquiry Word Count (LIWC), which was validated in [108]. Because LIWC was designed to

Table 4.3 – LIWC meta-categories, and examples of categories. N denotes the number of categories per meta-category.

Meta-category	N	Examples of categories
1. Linguistic processes	26	Word count, articles, verbs, adverbs
2. Psychological processes	32	
Social processes	3	Family, friends, humans
Affective processes	5	Negative emotions, positive emotions
Cognitive processes	8	Causation, tentative, exclusion
Perceptive processes	3	Sight, audition, touch
Biological processes	4	Body, health, sex
Relativity	3	Movement, space, time
3. Personal concerns	7	Work, leisure, money, death
4. Spoken categories	3	Consent, phatic, fillers
5. Punctuation	12	Commas, semi-columns, interrogation marks

process raw text data, no preprocessing on the transcript was required. All categories related to punctuation were dropped because they are not relevant to spoken speech. Categories which very seldom occured ($max < 1\%$) were also discarded. Highly skewed categories ($skew > 1$) were log-transformed ($x' = log(1+x)$, where x' and x denote the transformed and original features, respectively). Finally, all categories were standardized, such that each category had zero mean and unity variance. The values obtained for each LIWC category as feature vector (54 categories retained) represent the applicant's vocal content during the interview.

4.5 Feature subsets

Three types of behavioral cues were extracted and will be analyzed in the next chapter. The first feature subset is **NVB-CUES**, and consists of audio-visual nonverbal cues extracted for both the applicant and the interviewer. Feature subset **BODY-CUES** forms a representation of the applicant's body communicative behavior. Finally, the **VERBAL-CUES** feature subset consists of linguistic and paralinguistic categories based on manual transcriptions of the applicant's speech. Table 4.4 summarizes the feature subsets extracted in this chapter and used in the next chapters. It also highlights the variety of behavioral cues extracted and studied during the course of this thesis.

Table 4.4 – Feature subsets extracted from the SONVB interview dataset.

Feature subset	Description	Persons	Section
NVB-CUES	Audio-visual nonverbal cues	Applicant, interviewer	Section 4.2
BODY-CUES	Body communication cues	Applicant	Section 4.3
VERBAL-CUES	LIWC verbal cues	Applicant	Section 4.4

4.6 Conclusion

In this chapter, we presented our method to extract behavioral cues from employment interviews. The goal of this step was to provide an accurate representation (*i.e.*, a feature vector) of the interaction between the job applicant and the interviewer.

First, we developed a multimodal method to extract head nods in natural interactions, leveraging on the finding in social psychology stating that head gestures are conditioned by the speaking status [63]. Our work brought the novel angle of examining the audio self-context to improve the detection of head nods. We showed that using two different binary classifiers, depending on whether the person under analysis was speaking or not, could improve the detection accuracy of head nods. Additionally, the developed method allowed to detect subtle nods while keeping the number of false positives low.

Then, we extracted nonverbal cues from the audio and video modalities. To select what behavioral features to extract, we searched the psychology literature for cues consistently reported to play a role in job interviews. Then, we used available computational methods to extract behavioral features. Because the interviewer's behavior was shown to also affect interview outcomes [45], we not only extracted behavioral features for the applicant, but also for the interviewer. We then combined modalities and persons of interest to build multimodal and relational features.

To understand the role of an applicant's hand gestures and seated posture, we extracted body communication cues. Because of the difficulty to track body posture in seated configurations, and to benchmark the use of posture and gestures in an ideal case, we also used manual annotations of body communication, where classes were defined based on the psychology literature and their occurrence in the dataset. Additionally, we automatically extracted descriptors based on hand velocity as a raw representation of hand gestures. Because posture and gestures displayed by people considerably differ depending on whether they are speaking or silent [78], we included the speaking status in the feature extraction process.

Finally, to examine the role of verbal content in the formation of hirability impressions, we collected manual speech transcriptions and extracted linguistic and paralinguistic features using the Linguistic Inquiry Word Count (LIWC) [130]. LIWC is a software based on a word categorization system that relates linguistic and paralinguistic categories to psychological constructs, and each word from the transcript found in the dictionary increments one or several categories to which the word belongs.

Apart from the multimodal head nod detection method (Section 4.1), the goal of this chapter was *not* to formally evaluate the detection accuracy of the available computational tools used for the automatic extraction of behavioral features. However, cues were inspected manually to verify that all data was sane. Also, in this chapter we did not analyze the relationship between the extracted behavioral cues and the high-level interview social variables; this will be examined in detail in the next chapter.

5 Inference

One of the main objectives of this thesis is to assess the feasibility of automatically inferring applicant hirability during interviews. In this chapter, we present a computational framework for the automatic inference[1] of hirability in employment interviews. To this end, we used the SONVB interview dataset. To understand the relationship between nonverbal behavior and hirability impressions, we first completed a correlation analysis between audio-visual nonverbal cues (extracted for both the applicant and the interview) and hirability scores. Then, we investigated the use of standard machine learning techniques to predict hirability variables in a regression task. To our knowledge, our work is the first one focusing on the automated inference of employment interview outcomes from audio and visual nonverbal cues. We approach this problem from a behavioral, face-to-face perspective, where sensing, feature extraction, and social inference are (mostly) automated.

This chapter contains five main contributions. First, we evaluate a computational framework to infer the applicant's hirability based on the interaction during the interview. Second, we analyze the predictive validity of various feature groups (*e.g.* audio vs. visual cues, applicant vs. interviewer cues). Third, we compare the prediction performance obtained using psychometric questionnaire data as features with the one obtained using nonverbal cues. Fourth, we investigate the use of multimodal applicant body communication cues for the prediction of hirability impressions and personality traits. Fifth, we analyze the predictive validity of verbal content, represented by linguistic and paralinguistic categories. In this work, we demonstrate the feasibility of predicting hirability to some extent, achieving to explain up to 36.2% of the variance.

Figure 5.1 displays a graphic summary of our approach, in which the data and materials used for each section are specified. In Section 5.1, we first performed a correlation analysis between audio-visual nonverbal cues and the hirability scores. In Section 5.2, we defined the

[1] In computer science, the task of guessing the value of an unseen subject based on its feature representation is designated by the term *inference*, whereas the word *prediction* is used in psychology. In computer science, prediction designates the task of guessing the future value of a temporal sequence. However, in this thesis we use the two terms interchangeably, therefore the term *prediction* should be understood in the psychological sense.

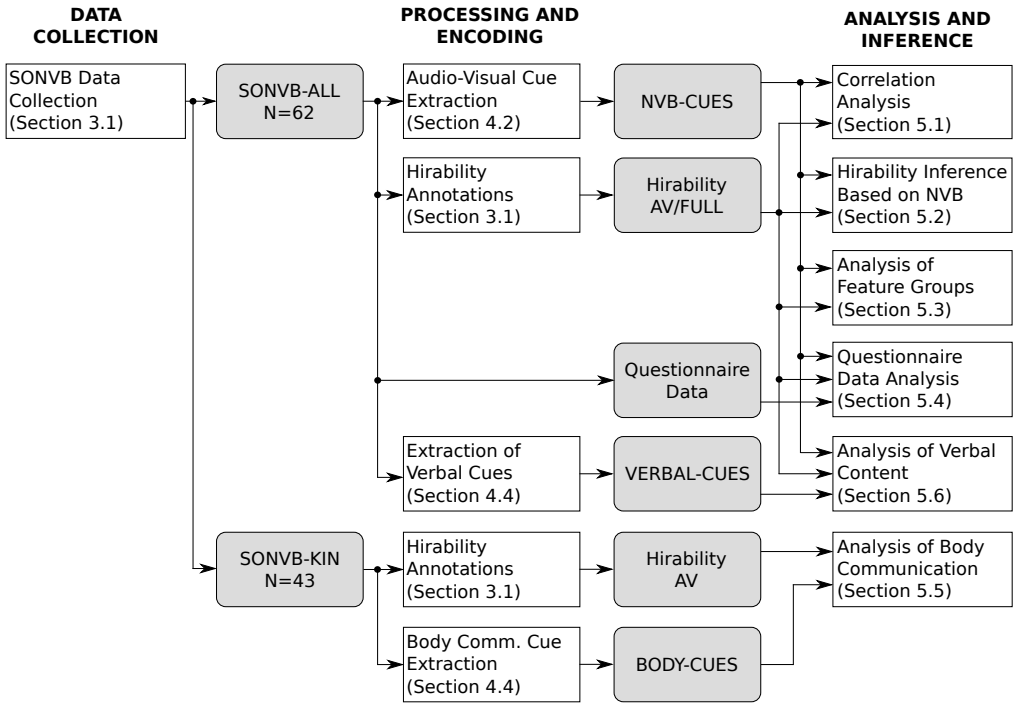


Figure 5.1 – Overview of Chapter 5.

inference task as a regression problem, where the goal was to predict the manually annotated hirability scores, and we evaluated several dimensionality reduction and regression methods for the inference of hirability scores from audio-visual nonverbal features. In Section 5.3, we compared the predictive power of feature groups, *e.g.* applicant *vs.* interviewer cues, and audio *vs.* visual cues, using ridge regression as inference method. We compared the predictive validity of questionnaire data with the performance obtained using nonverbal cues as predictors in Section 5.4. Then, in Section 5.5 we investigated the use of applicant body communication cues for the prediction of hirability and personality, and showed that conditioning on the speaking status improved the prediction performance. Last, in Section 5.6 we analyzed the predictive validity of Linguistic Inquiry Word Count (LIWC) as a representation of verbal content. Sections 5.1-5.4 were originally published in [97], while Section 5.5 originally appeared in [98].

5.1 Correlation analysis

In this section, we used the **SONVB-ALL** data subset (all 62 interviews) with the **AV** hirability annotation scheme (see Section 3.5 and Figure 5.1). As predictors, we used the **NVB-CUES** feature subset, presented in Section 4.2.

As a first step, we analyzed the linear relationships between the extracted behavioral cues and the hirability measures. Pearson's pairwise correlations between the extracted behavioral cues

and the hirability scores were computed. Nonverbal features which significantly correlated with hirability variables ($p < .05$) are reported in Table 5.1.

5.1.1 Applicant behavior

We first observe that applicant cues based on speaking activity and voiced rate statistics were consistently correlated with all hirability variables. Specifically, applicants who spoke longer, faster, had longer speaking turns, and required less number of turns to answer the questions obtained better hirability ratings than candidates who did not. This finding suggests that *fluency*, *i.e.* the ability to deliver a message quickly and clearly, played a role in the formation of hirability impressions. Also, applicants who spoke longer and had longer speaking turns were perceived as more hirable. Previous psychology studies [90, 45] have already suggested a relationship between applicant fluency and employment interview outcomes, therefore our findings are supported by previous work in psychology.

To a lesser degree, applicant face optical flow was found to be positively correlated with some hirability ratings (hiring decision and communication). Applicants who displayed more head motion received better hirability ratings. This observation finds some support in psychology literature [20, 57], where the amount of applicant head motion is positively correlated with interview outcomes. Similarly, applicant statistics on WMEIs (proxy for general visual motion) were found to be positively correlated with the hirability score of persuasion. This finding also goes along the lines of previous psychology literature suggesting a link between applicant kinetic expressiveness and job interview outcomes [20].

5.1.2 Interviewer behavior

One of the novelties of our work is the systematic study of the behavioral cues from the interviewer. Interviewer cues related to visual back-channeling, visual motion (head optical flow and some WMEI statistics), speaking activity, and prosody (voiced rate) were correlated with most hirability variables. In short, the interviewer spoke faster, had fewer speaking turns, produced more visual back-channels, and moved more in the presence of highly hirable candidates. This observation suggests that the interviewer's behavior was conditioned on the applicant: the interviewer acted differently whether she was in presence of a more or a less hirable job candidate. This possible instance of social mirroring (*a.k.a.* synchrony) between protagonists in an interaction is neither surprising nor new. An extensive body of literature (*e.g.*, [78]), has demonstrated the social influence a person can have on the other protagonists of an interaction in terms of nonverbal behavior. In the employment interview literature, researchers have studied the influence of the interviewer on the interview outcome by controlling his behavior (*e.g.* close *vs.* far distance, or cold *vs.* warm [83]), but to our knowledge previous work has not specifically studied the relationship between interviewer nonverbal cues and interview outcomes. A possible hypothesis to explain our findings is that the interviewer displayed some unconscious positive behavioral responses to highly hirable

Table 5.1 – Behavioral cues significantly correlated with at least one hirability score and corresponding Pearson's correlation coefficient ($p < .05$, $^{*}p < .01$, $^{\dagger}p < .005$). Not significantly correlated features are not reported. $N = 62$.

Cue	HirDecision	Communication	Persuasion	Conscience	StressRes
Applicant audio cues:					
Applicant # of short utterances	$-0.48^{\dagger}$	−0.30	−0.31	−0.26	
Applicant speaking time	$0.53^{\dagger}$		0.33^{*}	0.31	$0.44^{\dagger}$
Applicant # of turns	$-0.59^{\dagger}$	−0.31	$-0.40^{\dagger}$	−0.28	$-0.45^{\dagger}$
Applicant avg turn duration	$0.55^{\dagger}$	0.26	$0.41^{\dagger}$	0.33	$0.44^{\dagger}$
Applicant turn duration std	$0.41^{\dagger}$		0.33^{*}	$0.40^{\dagger}$	$0.38^{\dagger}$
Applicant max turn duration	$0.39^{\dagger}$		0.30	0.36^{*}	$0.39^{\dagger}$
Applicant fundamental frequency std	−0.30				
Applicant voiced rate avg	$0.57^{\dagger}$		$0.37^{\dagger}$		0.34^{*}
Applicant voiced rate std	0.30		0.30		
Applicant voiced rate median	$0.45^{\dagger}$		0.26		0.33
Applicant voiced rate lower quartile	$0.47^{\dagger}$	0.29			0.35^{*}
Applicant voiced rate upper quartile	$0.46^{\dagger}$		0.32		0.28
Applicant voiced rate max	0.33^{*}				$0.48^{\dagger}$
Applicant voiced rate entropy	$0.47^{\dagger}$		$0.39^{\dagger}$		0.32
Applicant visual cues:					
Applicant vertical optical flow avg	0.33	0.33			
Applicant vertical optical flow std		0.31			
Applicant WMEI avg				0.28	
Applicant WMEI std				0.26	
Applicant WMEI lower quartile				0.26	
Applicant WMEI entropy				0.26	
Applicant WMEI vertical center of mass				0.26	
Applicant coded expressiveness				$-0.38^{\dagger}$	−0.30
Interviewer audio cues:					
Interviewer # of short utterances	-0.34^{*}			-0.34^{*}	−0.28
Interviewer speaking time				−0.28	-0.34^{*}
Interviewer # of turns	-0.35^{*}			$-0.37^{\dagger}$	$-0.43^{\dagger}$
Interviewer avg turn duration	0.31	0.26		0.28	
Interviewer turn duration std	0.27	$0.43^{\dagger}$			$0.38^{\dagger}$
Interviewer max turn duration		$0.38^{\dagger}$			0.32
Interviewer voiced rate avg	0.29				
Interviewer voiced rate median	0.28				
Interviewer visual cues:					
Interviewer optical flow magnitude avg			0.30		
Interviewer vertical optical flow avg			0.33		
Interviewer # of nods		0.27			
Interviewer nodding time	0.33^{*}		$0.37^{\dagger}$		0.28
Interviewer # of visual BC	$0.40^{\dagger}$	0.31	0.34^{*}		0.35^{*}
Interviewer visual BC time	$0.45^{\dagger}$	0.29	$0.43^{\dagger}$		0.34^{*}
Interviewer WMEI std					0.31
Interviewer WMEI max					0.33^{*}
Interviewer audio-visual cues:					
Interviewer # of audio video BC ($\tau = 2000ms$)				−0.32	
Interviewer audio video BC time ($\tau = 2000ms$)				−0.28	
Interviewer # of nods while speaking				$-0.39^{\dagger}$	
Interviewer nodding while speaking time			−0.26	$-0.42^{\dagger}$	
Mutual cues:					
# of mutual short utterances ($\tau = 0ms$)	−0.30				
# of mutual short utterances ($\tau = 500ms$)	-0.35^{*}			−0.26	
# of mutual short utterances ($\tau = 1000ms$)	$-0.39^{\dagger}$		−0.26		
# of mutual short utterances ($\tau = 1000ms$)	$-0.40^{\dagger}$				
# of mutual short utterances ($\tau = 2000ms$)	−0.30				
Mutual short utterances time ($\tau = 500ms$)	$-0.40^{\dagger}$			−0.30	
Mutual short utterances time ($\tau = 1000ms$)	$-0.40^{\dagger}$			−0.31	
Mutual short utterances time ($\tau = 1500ms$)	$-0.38^{\dagger}$			−0.31	
Mutual short utterances time ($\tau = 2000ms$)	$-0.37^{\dagger}$			−0.28	

applicants, by producing more visual back-channels, speaking more fluently, and showing more visual motion.

5.1.3 Mutual cues

Mutual short utterances were negatively correlated with several hirability variables (hiring decision, persuasion, and conscience). They were the only relational cues connected to hirability measures. A possible explanation comes from the fact that these mutual short utterances were in practice short back-and-forth exchanges between the applicant and the interviewer. In most cases, the applicant would ask for a clarification on the question which was just posed, such as "In my private life?", to which the interviewer would answer "As you want". These short back-and-forth questions and answers were perceived negatively by the annotators: candidates answering questions at once without asking for clarifications received higher ratings. This finding could be related to applicant fluency: fluent candidates would answer the questions at once, without requiring further clarifications. This observation could also be related to nervousness as more nervous applicants would tend to hesitate more before answering the questions. Studying these hypotheses would require future work.

5.2 Inference of hirability variables based on nonverbal cues

In this section, we propose and evaluate a computational framework for the automatic inference of hirability in employment interviews. We defined the inference task as a regression problem, *i.e.* predicting the exact hirability scores, where each hirability variable was considered as an independent regression task. To this end, we used a two-step approach. The first step was dimensionality reduction, and the second was regression itself, where a regression model was trained and used to predict the hirability variables.

In this section, we used the **SONVB-ALL** data subset (all 62 interviews) with the **AV** hirability annotation scheme (see Section 3.5 and Figure 5.1). As predictors, we used the **NVB-CUES** feature subset, presented in Section 4.2.

5.2.1 Method

Dimensionality reduction

The goal of this step was to reduce the dimensionality of the behavioral feature vector. The feature dimensionality was not only high ($D > 140$) compared to the number of data points, it also contained by construction a large amount of redundant (*i.e.* highly inter-correlated features) and non-informative (*i.e.* cues independent of the hirability variables) data. Several standard dimensionality reduction methods were tested.

- **Low p-value features (pval)**. This method assumes that the relevant information is

contained in the features significantly correlated with the social variables. We selected features with $p < .05$.
- **Principal Component Analysis (PCA)**. PCA is a projection onto an orthogonal space of lower dimension. It learns the linear transformation such that the variance of the projected points is maximized [74]. In this study, the number of principal components was set such that 99.9% of the variance could be explained by the model.
- **All features (all)**. In order to test the improvement of the dimensionality reduction step, we also tested the case of taking all features as predictors for the regression step.

Please also note that several sequential feature selection methods [127] were tested, but due to the large dimensionality of the original feature set compared to the number of data points, the feature selection step was strongly over-fitting. Specifically, the feature selection methods yielded high validation results, but when tested within a cross-validation framework the performance drastically dropped due to the instability of the sets of selected features across folds. This problem is known in the machine learning community as *feature over-selection* [113]. For this reason, sequential feature selection methods were not used in this thesis.

Regression

In this step, the goal was to train a regression model for the prediction of the social variables. Several standard regression techniques were tested.

- **Ordinary least-squares (OLS)**. OLS minimizes the sum of squared errors between the observed and the predicted responses obtained using a linear model. It is the simplest regression model and is popular in psychology. The model assumes independent and identically distributed predictors, which in our study is the case only when PCA is used for dimensionality reduction.
- **Ridge regression (Ridge)**. Similarly to OLS, ridge regression minimizes the sum of squared errors between the observed and predicted responses of a linear model, but a regularization term is added to the cost function, which multiplies the l_2-norm of the regression coefficients. This regression penalty has the effect of shrinking the estimates towards zero, preventing the model to over fit [66].
- **Random forest (RF)**. Used for classification and regression, RF is based on the bootstrap aggregation of a large number of decision trees. In the regression case, standard decision trees split the feature space into hyper-cubic regions assigned to values [34]. RF aggregates the output of each separate decision tree by taking the average predicted value. RF has the advantage of being robust to over fitting and of not making strong assumptions about the input features.

We used a leave-one-interview-out cross-validation approach for training and testing the regression models. This framework used all but one interview for training, and kept the remaining one for testing. Model parameters were estimated using a 10-fold inner cross-validation approach.

5.2.2 Evaluation measures

We measured the performance of the automatic prediction models using the root-mean-square error ($RMSE$) and the coefficient of determination (R^2), as these are two widely used measures in psychology and social computing. As the baseline regression model, we took the average hirability score as the predicted value. The $RMSE$ is defined in Equation 5.1, where y_{gt} are the ground truth observed variables, y_{pred} are the predicted values, and N is the number of data samples:

$$RMSE = \sqrt{\frac{\sum(y_{gt} - y_{pred})^2}{N}}. \tag{5.1}$$

The coefficient of determination R^2 is based on the ratio between the mean squared errors of the predicted values obtained using a regression model and the baseline-average model. It is defined in Equation 5.2, where y_{gt} and $\bar{y}_{gt}$ are the observed variables and their mean; and y_{pred} are the predicted values. R^2 can be seen as the relative improvement over the baseline-average model. Note that negative value can be obtained when the evaluated model under-performs with respect to the baseline-average model.

$$R^2 = 1 - \frac{\sum(y_{gt} - y_{pred})^2}{\sum(y_{gt} - \bar{y}_{gt})^2} \tag{5.2}$$

Finally, significance levels were computed using Student's t-test on the difference between the squared residuals (*i.e.*, the difference between the predicted value and the ground truth score) of the tested regression model and the baseline-average model. The null hypothesis was defined as the mean being zero, assuming a Gaussian distribution and unknown variance. Cases where squared residuals have low average but high variance can result in low $RMSE$ and high R^2, but high p-values (*i.e.*, low significance levels).

5.2.3 Results

Table 5.2 shows the performance of the different models for the inference of hirability variables. Performance values for OLS regression were not reported as the method consistently performed worse than the baseline-average model, due to over fitting.

Results obtained for the hiring decision variable were significantly better than the baseline-average model for ridge regression ($p < .05$) independently of the dimensionality reduction technique, and for random forest using all nonverbal features. The best prediction result for hiring decision was obtained using ridge regression with all features and ridge regression

Table 5.2 – Performance (R^2 and $RMSE$) for the inference of hirability scores using different dimensionality reduction and regression methods ($^* p < 0.1$, $^\dagger p < 0.05$ for $RMSE$). $N = 62$.

	HirDecision		Communication		Persuasion		Conscience		StressRes	
Method	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$
Baseline-Avg	0.00	1.80	0.00	0.96	0.00	1.04	0.00	0.92	0.00	0.79
All-Ridge	0.36	$1.44^\dagger$	-0.07	0.99	0.08	1.00	-0.10	0.96	0.12	0.74
All-RF	0.27	$1.53^\dagger$	0.00	0.96	0.12	0.97^*	0.04	0.90	0.08	0.76
Pval-Ridge	0.33	$1.47^\dagger$	-0.05	0.98	0.11	0.98	-0.010	0.92	0.27	0.68^*
Pval-RF	0.29	1.52^*	0.02	0.95	0.05	1.01	-0.04	0.94	0.21	0.71^*
PCA-Ridge	0.36	$1.44^\dagger$	-0.07	0.99	0.07	1.00	-0.09	0.96	0.13	0.74
PCA-RF	0.08	1.73	-0.08	0.99	0.05	1.01	0.02	0.91	0.06	0.77

with PCA as dimensionality reduction ($R^2 = 0.36$ for both). Hiring decision prediction results obtained with random forest were significantly more accurate than the baseline-average model when no dimensionality reduction was applied prior to the regression step ($R^2 = 0.27$), and marginally significant using low p-value features as predictors ($R^2 = 0.29$).

For the variable of stress resistance, random forest and ridge regression using low p-value features as predictors produced marginally significantly better results than the baseline-average model (respectively, $R^2 = 0.27$ and $R^2 = 0.21$). Although not statistically significant, ridge regression with other dimensionality reduction methods yielded positive results ($R^2 = 0.12$ for all features and $R^2 = 0.13$ for PCA). For persuasion, the results obtained with random forest and all features were marginally more accurate than the baseline-average model ($R^2 = 0.12$). For the remaining hirability variables (communication and conscience), no method was able to outperform the baseline average model.

5.2.4 Discussion

The results found here show the feasibility of automatically inferring the hiring decision score to some degree. Moreover, the use of nonverbal behavioral features as a basis for predicting hirability is a valid hypothesis. The variable of stress resistance was also possible to predict, even if the results were only marginally more accurate than the baseline-average model. In contrast, the variables of communication, persuasion, and conscience were more difficult to infer. A possible hypothesis to explain this finding is that raters did not form their opinion from the cues which were extracted. Raters might rather have used more verbal content as a basis to form their opinion on these constructs than for the hiring decision. This hypothesis is investigated in Section 5.6.

To contextualize the achieved performance, we refer to existing work in psychology. In [123], Schmidt obtained $R^2 = 0.18$ from predictors composed of nonverbal cues and a variety of "meta-behaviors" such as attentiveness, empathy, or dominance. Gifford *et al.* [57] obtained R^2 values ranging from 0.49 to 0.62 for the prediction of motivation and social skills (both perceived and self-rated), which are slightly different constructs compared to hirability. A notable exception in the literature is the work by Parsons *et al.* [104] who reported $R^2 = 0.72$; the

authors of the paper themselves were surprised by this extremely high result and hypothesized that it could be an effect of the way hirability scores and nonverbal cues were annotated. In all these works, R^2 results were obtained using OLS regression without separating the data into training and test sets. From this standpoint, the performance results achieved here are comparable to the ones reported in the psychology literature, with the advantage that they were obtained using a rigorous prediction framework in the machine learning sense.

In terms of regression methods, ridge regression was the best-suited technique from the pool of methods tested for our task. The reason behind this finding may come from the fact that linear relationships between the features and the hiring decision exist, as suggested by the statistical analysis performed in Section 5.1; in other words, the linear assumption used in ridge regression likely held. Another interesting finding is that dimensionality reduction did not improve the prediction of ridge regression for the hiring decision score. This suggests that ridge regression was able to find the informative patterns without needing a pre-processing step, with the l_2-regularization term implicitly selecting the most informative features by assigning low weights to redundant or uninformative features. For stress resistance, low p-value dimensionality reduction improved the accuracy, suggesting that the informative data was contained in significantly correlated features. PCA with ridge regression produced results of similar accuracy, compared with the ones obtained with no dimensionality reduction. This suggests that the transformation retained the informative data, but did not result in more informative patterns. Also, this suggests that although the independence assumption in the predictors was not held, ridge regression was still able to produce good predictions. On the other hand, PCA coupled with random forest showed poor prediction performance, which was not the case for the other dimensionality reduction techniques.

5.3 Analysis of Feature Groups

In this section, we analyze the predictive power of feature groups. Feature groups were defined based on the person and the modality from whom the features were extracted. We used the **SONVB-ALL** data subset (all 62 interviews) with the **AV** hirability annotation scheme (see Section 3.5 and Figure 5.1). As predictors, we used the **NVB-CUES** feature subset, presented in Section 4.2.

5.3.1 Method

Four groups were defined, based on the protagonist from whom the nonverbal cues were extracted: applicant, interviewer, mutual, and all. For each person-related group, the features were further separated into three sub-groups based on the modality: audio, video, and all. Based on the results obtained in Section 5.2 showing that ridge regression with no dimensionality reduction produced in most cases the best prediction performance, the analysis of feature groups was performed using this inference method. Please note that this analysis was also done for random forest with no dimensionality reduction and yielded similar results,

therefore the results were not reported. We used leave-one-interview-out cross-validation to train and test the inference method, and 10-fold inner cross-validation to select the best ridge parameter, as in Section 5.2.

5.3.2 Results

The results obtained using the different feature groups as predictors are reported in Table 5.3. Feature groups who yielded the best prediction results were audio cues extracted from both the applicant and the interviewer ($R^2 = 0.40$), interviewer visual cues ($R^2 = 0.37$), applicant audio cues ($R^2 = 0.32$), applicant audio and visual cues ($R^2 = 0.25$), and interviewer audio and visual cues ($R^2 = 0.22$).

Applicant features were predictive of the hiring decision (all applicant features: $R^2 = 0.25$, $p < .1$); results showed that the predictive applicant features stemmed from audio (applicant audio features: $R^2 = 0.32$, $p < .05$) and not from video (negative R^2). In the light of the statistical analysis conducted in Section 5.1, these results are not surprising as only one visual feature was found to be significantly correlated with the hiring decision. However, the result goes against related work in psychology suggesting a relationship between hirability and several visual cues (gaze, smiles, head gestures, or physical attractiveness [20]). In our case, adding the applicant visual features to the applicant audio features did not improve the prediction accuracy; rather, it decreased the performance. Note however that our features did not include gaze.

Interestingly, the results obtained for the interviewer group showed good performance. In this case, interviewer visual cues showed the best accuracy ($R^2 = 0.37$, $p < .1$), whereas interviewer audio cues were not predictive. Combining audio and visual interviewer cues decreased the performance ($R^2 = 0.22$, $p < .1$) compared to the visual cues taken alone.

When grouped together, audio features extracted from the two protagonists showed the best performance across all feature groups ($R^2 = 0.40$, $p < .05$). Adding the interviewer audio cues to the applicant audio cues increased the prediction accuracy (from $R^2 = 0.32$ to $R^2 = 0.40$), even if the interviewer audio cues and the mutual audio cues were not predictive in isolation. For the group of visual cues extracted from both protagonists, the results suggest an opposite tendency: interviewer visual cues showed good accuracy ($R^2 = 0.37$), but adding the non-predictive applicant visual cues decreased the accuracy dramatically ($R^2 = 0.01$).

Also note that the significance levels are not directly linked to the R^2 and $RMSE$ values. Indeed, the applicant-audio group has lower R^2 and higher $RMSE$ ($R^2 = 0.32$, $RMSE = 1.49$) than interviewer-video ($R^2 = 0.37$, $RMSE = 1.42$), but higher significance level. These results may seem conflicting, but can be explained by the fact that the squared residuals for applicant-audio had lower variance but higher average than for interviewer-video, resulting in a lower p-value.

5.3.3 Discussion

Applicant cues were predictive of the hiring decision score to some degree. More specifically, the relevant information stemmed from the audio modality, whereas visual features produced low prediction results. The combination of audio and visual cues decreased the performance compared to audio cues taken in isolation. One hypothesis to explain why applicant visual cues were not predictive could be that raters could have used visual features which were not extracted in this study, such as applicant body posture, or fine-grain gaze patterns. The systematic examination of this hypothesis will be the subject of future work.

Interviewer cues were predictive of the hiring decision, which is in our opinion an interesting finding: one can to some extent infer the hirability of an applicant by observing the interviewer only. In other words, the behavioral responses of the interviewer were valid predictors for the hiring decision variable, which suggests an instance of synchrony occurring during the interaction. Previous work in psychology has established that nonverbal synchrony could be associated with different social outcomes, such as the quality of the patient-therapist relationship [111]. Here, the predictive validity of interviewer cues stemmed from the video modality. From Table 5.1, the features of interest were related to interviewer visual back-channeling. Combining audio cues to visual cues however decreased the prediction performance. This finding implies that by only looking at the interviewer, it is possible to make inferences on the applicant; this approach (looking at others for inferring things about self) has been used in one previous computational study in a different setting [106]. The study of nonverbal synchrony is a subject in itself and would deserve a larger treatment as part of future work [46].

When considering modalities without taking the person of interest into account, audio cues showed the best prediction performance. Interestingly, combining applicant audio cues (high predictive validity) and interviewer audio cues (low predictive validity) actually improved the prediction performance. This finding suggests that interviewer audio cues contained some informative data, but were only useful when combined to applicant audio cues. For the visual modality, the results showed an opposite trend: combining applicant visual cues (low predictive validity) with interviewer visual cues (high predictive validity) dramatically decreased the prediction performance. These results are interesting from the multimodal processing standpoint.

5.4 Analysis of Questionnaire Data

The use of psychometric questionnaires for the personnel selection process is a common practice in human resources. Questionnaires are used to assess social constructs related to the task at hand. Psychology researchers have identified a number of social constructs frequently assessed during job interviews, such as intelligence, knowledge and skills, personality traits, applied social skills, interests and preferences, organizational fit, and physical attributes [69]. In this section, we analyze the predictive validity of psychometric questionnaires, in relation with hirability scores.

Table 5.3 – Performance (R^2 and $RMSE$) for the prediction of the hiring decision score using different feature groups as predictors, and using ridge regression with no dimensionality reduction as inference method ($^{*}\,p < 0.1$, $^{\dagger}\,p < 0.05$). $N = 62$.

	HirDecision	
Feature Group	R^2	$RMSE$
Baseline-Avg	0.00	1.80
Applicant-Audio	0.32	$1.49^{\dagger}$
Applicant-Video	-0.05	1.84
Applicant-All	0.25	1.55^{*}
Interviewer-Audio	0.03	1.77
Interviewer-Video	0.37	1.42^{*}
Interviewer-All	0.22	1.59^{*}
Mutual-Audio	0.05	1.76
Mutual-Video	-0.03	1.83
Mutual-All	0.05	1.75
All-Audio	0.40	$1.39^{\dagger}$
All-Video	0.01	1.79
All	0.36	$1.44^{\dagger}$

Table 5.4 – Pairwise correlations between questionnaire data and hirability scores ($^{*}\,p < .05$, $^{\dagger}\,p < .01$). $N = 62$.

	HirDecision	Communication	Persuasion	Conscience	StressRes
Extraversion	$0.42^{\dagger}$	0.27^{*}	0.20	0.27^{*}	0.27^{*}
Openness	-0.04	-0.16	0.02	-0.05	0.06
Neuroticism	-0.26^{*}	-0.11	-0.17	-0.26^{*}	-0.22
Agreeableness	-0.01	-0.07	-0.13	-0.07	-0.04
Conscientiousness	0.05	-0.14	0.07	0.23	0.12
Communication skills	0.09	-0.01	0.05	0.14	0.16
Persuasion skills	-0.01	-0.05	-0.05	0.01	0.10
Intelligence	0.08	0.19	0.07	-0.14	-0.05

To analyze the use of questionnaires for hirability prediction, it was necessary to assume that the recruiter also had access to the questionnaire data. To this end, we used the **SONVB-ALL** data subset (all 62 interviews) with the **FULL** hirability annotation scheme (see Section 3.5 and Figure 5.1). As predictors, we used the **NVB-CUES** feature subset, presented in Section 4.2.

5.4.1 Correlation analysis

Pairwise correlations between the questionnaire variables (Section 3.3) and the hirability scores are reported in Table 5.4. We observe that extraversion was correlated with all hirability variables except persuasion. This finding is supported by the psychology literature showing a relationship between extraversion and performance in jobs characterized by a high level of social interactions, such as in sales, marketing, or management [22]. Openness to experience

Table 5.5 – Performance (R^2 and $RMSE$) for the inference of hirability scores using questionnaire data as predictors and ridge regression with no dimensionality reduction ($^{*}p < .1$, $^{\dagger}p < .05$). Results are then compared to results obtained using nonverbal cues as features.

	HirDecision		**Communication**		**Persuasion**		**Conscience**		**StressRes**	
	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$
Baseline-Avg	0.00	0.89	0.00	0.89	0.00	0.97	0.00	0.86	0.00	0.78
Personality	0.04	0.87	-0.04	0.91	-0.09	1.00	0.00	0.85	-0.06	0.79
Communication	-0.03	0.90	-0.03	0.90	-0.03	0.97	-0.07	0.88	-0.06	0.79
Intelligence	-0.04	0.90	-0.04	0.91	-0.04	0.98	-0.06	0.88	-0.03	0.78
All-Quest	-0.03	0.90	-0.03	0.90	-0.07	0.99	-0.03	0.86	-0.06	0.79
NVB	0.29	$0.75^{\dagger}$	-0.08	0.93	0.02	0.96	0.19	0.77	0.01	0.78
NVB+Quest	0.29	$0.75^{\dagger}$	-0.09	0.93	0.01	0.96	0.19	0.77	-0.01	0.78

was not correlated with any of the hirability variables. Psychology research has also found no strong relationship between this trait and performance or interview ratings [117]. Neuroticism was negatively correlated with the scores of hiring decision and conscience, which also goes along the lines of related work suggesting that this characteristic negatively affects the employability of candidates [117]. No significant correlation was found for the agreeableness trait, which does not contradict the previous work in psychology as this trait was found to be related with performance only for certain occupations, such as team-work or customer service [117], which was not the type of job for which candidates were applying. Similarly, no significant correlation was found for the conscientiousness trait. This finding, however, is surprising as psychology literature showed a significant relationship between this trait and job performance across all types of occupations [22]. Also, although one of the hirability scores (conscience) was specifically targeted at assessing this trait, the pair-wise correlation between the two variables was only marginal ($p < .1$). The questionnaire variable related to communication skills did not have any significant correlation with the hirability variables, even if one of them (communication) was targeted at assessing it. The same observation can be made for the questionnaire variable of persuasion. Finally, intelligence was found to have no significant correlation with the hirability scores. This observation does not match the psychology research which consistently showed a significant relationship between general intelligence and job performance across multiple types of occupations [69].

5.4.2 Prediction

To analyze the predictive validity of psychometric questionnaires with respect to hirability, we used the regression task introduced in Sections 5.2 and 5.3.We used ridge regression with no dimensionality reduction, as it consistently produced the most accurate prediction in Section 5.2. Furthermore, we separated questionnaire data into three groups depending on the social construct they belonged to: personality traits, communication and persuasion skills, and intelligence. Finally, we compared the results with the ones obtained using nonverbal features, and the combination of questionnaire data with nonverbal behavior.

Results are reported in Table 5.5. The prediction results achieved using questionnaire data

as features were less accurate than the baseline-average model. Combining questionnaire data and nonverbal behavior did not improve the prediction accuracy compared to taking nonverbal features alone ($R^2 = 0.29$ for both). Experiments using random forest were also conducted and the results obtained were similar to the ones produced with ridge regression, therefore they were not reported here.

5.4.3 Discussion

Questionnaire data held no predictive validity for the inference of hirability variables. Even if two personality traits were found to be significantly correlated with hiring decision, they were not useful for predicting the hirability scores. In comparison, nonverbal cues produced prediction results significantly more accurate than the baseline. The use of questionnaire data put in combination with behavioral features did not improve the prediction accuracy. These findings suggest that raters used nonverbal behavior rather than questionnaire data as basis to form their opinion about the applicants' hirability. In other words, not only was nonverbal behavior more useful than questionnaire data for the prediction of the hiring decision score, but questionnaire data provided no information for inferring hirability.

Given the broad use of questionnaires in employment interviews, these results are surprising at first glance. Indeed, the results seem to contradict previous psychology research showing the validity of certain constructs such as intelligence or personality in the personnel selection process [69, 117]. Previous psychology studies have used personality traits as predictors for the regression of hirability or similar constructs and have reported results ranging from $R^2 = 0.16$ [35] to $R^2 = 0.43$ [41] using OLS regression. The results obtained in these works were however not obtained from a prediction task in the machine learning sense, *i.e.* there was no separation between training and test sets. We were able to reproduce results similar to [35] and [41] using the same approach (OLS regression with no cross-validation), obtaining $R^2 = 0.23$ for the hiring decision score using the Big-Five traits as independent variables, and similar results for the other hirability variables (R^2 ranging from 0.11 to 0.18). However, when separating training and test sets, we observed a drastic performance drop. This observation highlights the necessity to separate training and test sets to assess the predictive power of the independent variables of interest. It also shows that obtaining significant correlations is necessary, but not sufficient to have a reliable prediction model.

5.5 Multimodal analysis of body communication cues

In this section, our objective is to analyze the role of the applicant's body communicative behavior in the formation of hirability impressions. More specifically, we investigate whether hirability impressions and self-rated personality can be predicted using body communication cues alone; moreover, leveraging on the multimodal nature of posture and gestures, we examine whether the knowledge of the speaking status can be used to improve the prediction of personality and hirability.

To address these research questions, several tasks were defined. First, we used the **SONVB-KIN** data subset (Section 3.5), from which we extracted the applicant body communication features presented in Section 4.3 (the **BODY-CUES** feature subset), including the speaking status in the extraction process. As a next step, we analyzed the differences of body communication cues depending on whether the applicant was speaking or silent. Last, we evaluated the predictive validity of the extracted nonverbal cues with respect to hirability impressions obtained using the **AV** annotation scheme (see Section 3.5) and self-rated personality using a regression task.

The main contributions of this section are: (1) the systematic analysis of body communication cues for the prediction of social constructs, and (2) the exploitation of the multimodal nature of body communication to improve the prediction performance of personality and hirability. Thus, we extend the previous sections of this chapter by predicting personality traits in addition to hirability, as well as by focusing on postures and gestures. This work was originally published in [98] and was a collaboration with Alvaro Marcos Ramiro (PhD student at University of Alcala, Spain).

5.5.1 Analysis of speaking status

In order to test whether our initial assumption stating that body communication was conditioned on the applicant's speaking status, we computed the Student's t-test to examine whether significant differences in feature values between speaking and silent existed. In Table 5.6, we display the significantly different features ($p < 0.05$), and report whether the larger value was associated with moments when the job applicant was silent or speaking.

We observe that job applicants gestured more when they were speaking (more *gestures* and *gestures on table* time, longer events, larger range of durations; higher hand speeds; higher hand accelerations). Inversely, interviewees self-touched and kept their hands on the table longer when listening to the interviewer. This findings validate our main assumption of the multimodal nature of hand gestures and body posture, based on the nonverbal communication literature [78, 92]. Furthermore, we observe that the automatic features based on hand speed and hand acceleration were also conditioned on the speaking status.

5.5.2 Prediction of hirability and personality

Method

In order to analyze the predictive validity of body posture with respect to self-rated personality traits and hirability impressions, we defined a regression problem which aims at predicting the exact hirability and personality scores, where each social variable is considered as an independent regression task. To this end, we used a leave-one-interview-out cross-validation strategy. Two regression methods were used for predicting personality and hirability, namely ridge regression and random forest.

Table 5.6 – Feature significantly different ($p < .05$) between speaking and silent, using Student's t-test ($N = 43$).

Feature group	Larger feature value for silent	Larger feature value for speaking
Hidden Hands		# of events
Self-Touch	rel. time, median, max., min., quartiles, range	# of events
Hands on Table	rel. time, mean, std, upper quartile	# of events
Gestures		rel. time, mean, median, std, max., upper quartile, range, exist, # of events
Gestures on Table		rel. time, mean, median, std, max., min., range, quartiles, # of events, exist
Hand Speed	min, zero-crossing rate	mean, median, max., quartiles, range, non-zero prop.
Hand Acceleration		mean, median, max., min., quartiles, non-zero prop.

Given the large number of features ($D > 300$) compared to the number of data points ($N = 43$), we decided to analyze sub-groups of features independently. This allowed the regression model to be correctly learned, and enabled the analysis of the predictive validity of specific postures and speaking cases. For the nonverbal features based on the manual annotations, five feature groups were defined based on the annotated body activity classes (described in Table 4.1). For the automatic cues, we used the hand movement e_H, the hand acceleration a_H, the activity histogram $\mathbf{h}_{OF}$, and the acceleration histogram $\mathbf{a}_{OF}$ cues as four feature groups.

In order to test whether exploiting the speaking status improves the prediction accuracy, we further segmented the feature groups into four sub-groups: (1) unimodal features, *i.e.* obtained without taking into account the speaking status, (2) silent features only, (3) speaking features only, and (4) aggregated features, *i.e.* the concatenation of unimodal, silent, and speaking cues. To evaluate the prediction accuracy of our method, we used the coefficient of determination R^2.

Results and discussion

In Table 5.7, we report the results for which the R^2 values were higher than 0.1. From those findings, several observations can be made. Except for communication ability, all hirability scores were inferred above the R^2 threshold using multimodal body communication cues. Importantly, we achieved $R^2 = 0.21$ for the hiring decision score using automatically extracted activity histogram features (aggregation of unimodal, speaking, and silent) and ridge regression as a prediction method. This finding demonstrates the potential of predicting job interview outcomes using body communication cues.

For personality, we show that prediction can be done using body communication cues only,

Table 5.7 – Prediction results for hirability impressions (1-5) and self-rated personality (6-10) using manual (M) and automatic (A) cues. R^2 was used to evaluate the prediction performance. Only results with $R^2 > 0.1$ are reported. $N = 43$.

Hirab. variable	Feature group	Spking status	Regr. method	R^2
1. HirDec	Gesturing (M)	Silent	Ridge	0.20
	Activ. hist. (A)	Silent	Ridge	0.18
	Activ. hist. (A)	Silent	RF	0.14
	Activ. hist. (A)	Aggr.	Ridge	**0.21**
	Activ. hist. (A)	Aggr.	RF	0.14
	Acc. hist. (A)	Silent	Ridge	0.18
	Acc. hist. (A)	Silent	RF	0.11
2. Comm	-	-	-	-
3. Consc	Hid. hnds (M)	Speak	RF	0.17
	Hid. hnds (M)	Aggr.	RF	0.10
	Gest. table (M)	Speak	RF	0.16
	Gest. table (M)	Aggr.	Ridge	0.15
	Gest. table (M)	Aggr.	RF	**0.20**
	Activ. hist. (A)	Silent	RF	0.12
	Activ. hist. (A)	Aggr.	RF	0.11
4. Persuas	Hid. hnds (M)	Aggr.	Ridge	**0.24**
	Activ. hist. (A)	Silent	RF	0.20
	Activ. hist. (A)	Aggr.	RF	0.11
	Activ. hist. (A)	Aggr.	Ridge	0.13
	Acc. hist. (A)	Silent	Ridge	0.14
	Acc. hist. (A)	Silent	Ridge	0.11
5. StrRes	Activ. hist. (A)	Aggr.	RF	**0.10**

Pers. variable	Feature group	Spking status	Regr. method	R^2
6. Extra	Hid. hnds (M)	Unimod.	RF	**0.17**
	Hid. hnds (M)	Speak	RF	0.12
	Hid. hnds (M)	Aggr.	Ridge	0.15
	Hid. hnds (M)	Aggr.	RF	0.14
	Slf-tch (M)	Unimod.	RF	0.11
	Slf-tch (M)	Aggr.	RF	0.13
	Gest. table (M)	Unimod.	RF	0.15
	Gest. table (M)	Speak	RF	0.14
7. Open	Slf-tch (M)	Unimod.	Ridge	0.11
	Slf-tch (M)	Silent	RF	0.10
	Hnds table (M)	Silent	RF	**0.24**
	Hnds table (M)	Silent	RF	0.18
8. Neuro	-	-	-	-
9. Agree	Hid. hnds (M)	Speak	RF	**0.14**
	Gest. table (M)	Speak	RF	0.11
10. Consc	Hid. hnds (M)	Unimod.	Ridge	0.14
	Hid. hnds (M)	Silent	Ridge	**0.17**
	Gest. table (M)	Unimod.	Ridge	0.14
	Gest. table (M)	Silent	Ridge	**0.17**

which to our knowledge had not been analyzed systematically prior to this work. Using such nonverbal features showed prediction performance comparable to the related work in social computing. Extraversion prediction ($R^2 = 0.17$) was found to be less accurate than in the state of the art (*e.g.* [29]), but results obtained for openness to experience ($R^2 = 0.24$), agreeableness ($R^2 = 0.14$), and conscientiousness ($R^2 = 0.17$) were positive.

Only six prediction scores where $R^2 > 0.1$ were achieved using unimodal features (*i.e.* without taking into account the speaking status of the job applicant when extracting the cues). In comparison, 33 prediction scores were achieved by using body communication cues conditioned on the speaking status. This finding further shows the intimate link between speaking status and body communication in this job interview setting. Furthermore, we show that leveraging on this finding can improve the prediction of social constructs.

We observe that automatic hand activity cues were predictive of hirability ratings. Indeed, the best prediction results for the hiring decision and stress resistance were achieved using automatic cues based on activity histograms. For the hirability variables of persuasion and conscience, the use of automatic features decreased the prediction accuracy compared to manual features (from 0.24 to 0.21 and 0.20 to 0.12, respectively). This finding suggests that manual annotations of postures might not be necessary, depending on the social construct of interest. The use of automatic body communication cues was however found to show poor performance for self-rated personality traits.

5.6 Analysis of verbal content

In this section, our objective is to analyze the role of verbal content in the formation of hirability impressions during job interviews. To complement Sections 5.1-5.5, which focused on nonverbal behavior, here we investigate whether language style, represented by the Linguistic Inquiry Word Count (LIWC), a combination of linguistic and paralinguistic categories extracted from manual interview transcripts, can be predictive of hirability impressions. Additionally, we take a close look on what categories are associated with high hirability ratings. To our knowledge, no previous computational work has focused on verbal content in job interviews. This work is relevant as it provides additional insights on the way hirability impressions are formed, in complement to the analyses conducted on nonverbal behavior.

To this end, we used the full SONVB interview dataset (**SONVB-ALL**), with hirability impressions gathered using the **AV** scheme (see Section 3.5). As predictors, we used the verbal features presented in Section 4.4 (the **VERBAL-CUES** feature subset). This section is structured as follows. In Section 5.6.1, we conducted a correlation analysis between LIWC categories and hirability ratings. Then, we evaluated the use of LIWC features for the prediction of hirability impressions in Section 5.6.2. We combined LIWC categories with applicant and interviewer nonverbal cues and evaluated the prediction results in Section 5.6.3. We finally discuss the results in Section 5.6.4.

5.6.1 Correlation analysis

First, we analyzed the linear relationships between verbal features and hirability impressions. Pearson's pairwise correlations between LIWC categories and hirability scores are shown in Table 5.8; only categories which were significantly correlated ($p < .05$) with at least one hirability variable are displayed.

We first observe that some verbal cues associated to linguistic processes were consistently and significantly correlated with most hirability variables. Specifically, applicants who used more words and a larger amount of words per sentence were perceived as more hirable. According to [38], these categories are associated to talkativeness and verbal fluency; therefore this observation aligns with the findings of Section 5.1 and previous psychology studies [90, 45], stating that fluency was related to hirability impressions. In fact, word count (WC) and words per sentence (WPS) could be seen as nonverbal cues, because they account for the amount of speech rather than the content itself; this observation is corroborated by the fact that these two cues were strongly correlated with the applicant speaking time ($r_{WC} = 0.67$ and $r_{WPS} = 0.58$) and applicant average turn duration ($r_{WC} = 0.58$ and $r_{WPS} = 0.53$).

Another observation is that verbal categories related to affect were found to be negatively correlated with most hirability impressions. Interestingly, both negative and positive emotions were negatively correlated with perceptions by raters. To a lesser degree, the present tense category (associated with "living the here and now" [38]) negatively correlated with persuasion; and word>6 letters (associated with social class, education [38]) positively correlated with persuasion, suggesting that applicants who displayed more verbal immediacy and more emotions were perceived as less hirable.

In terms of cue utilization, we observe that stress resistance and the hiring decision score had respectively 8 and 7 significantly correlated verbal features, with some cues having large correlation coefficients (up to 0.56 for hiring decision and 0.45 for stress resistance). This finding is encouraging for the task of hirability prediction based on verbal behavioral features, suggesting that this representation of verbal content is appropriate. For persuasion and conscientiousness, although the number of significantly correlated features was similar (resp. 7 and 6), correlation values were lower. For communication, cue utilization was low (only 2 significant features), which suggests that the connections were weaker.

5.6.2 Prediction of hirability based on linguistic categories

To assess the validity of verbal cues for the task of predicting hirability variables, we completed inference experiments, where the task was defined as a regression problem. We used ridge regression with no dimensionality reduction as this method was shown to obtain accurate results in Section 5.2. We used a leave-one-interview-out cross-validation strategy as in previous cases. For evaluation, we used the coefficient of determination (R^2) and the root-mean-square error (RMSE). Prediction results are displayed in Table 5.9.

Table 5.8 – Linguistic Inquiry Word Count (LIWC) categories significantly correlated with at least one hirability variable, using Pearson's correlation coefficient ($p < .05$, $^{*}p < .01$, $^{\dagger}p < .005$). $N = 62$.

LIWC category	HirComm.	HirPers.	HirConsc.	HirStressRes.	HirDecision
Linguistic processes:					
Word count		0.29	0.34^{*}	$0.37^{\dagger}$	$0.37^{\dagger}$
Words/sentence	$0.38^{\dagger}$	0.31	0.35^{*}	$0.45^{\dagger}$	$0.56^{\dagger}$
Words>6 letters		0.31			
Dictionary	−0.30				
You				−0.28	$-0.36^{\dagger}$
Present tense		−0.26			
Conjunction				0.29	0.27
Psychological processes:					
Affect		$-0.38^{\dagger}$	−0.26	$-0.41^{\dagger}$	$-0.42^{\dagger}$
Positive emotions		$-0.39^{\dagger}$		$-0.38^{\dagger}$	$-0.37^{\dagger}$
Negative emotions					−0.27
Inclusion				0.34^{*}	
See			0.26		
Biologic processes			−0.28		
Health			−0.29	−0.27	
Personal concerns:					
Work		0.35^{*}			

Prediction results obtained with all LIWC features for hiring decision and stress resistance were observed to be slightly above the baseline average model, suggesting that the information contained in verbal behavior was not used much by raters in the formation of hirability impressions. However, when using the categories of word count (WC) and words per sentence (WPS) alone, accuracies of $R^2 = 0.28$ and $R^2 = 0.16$ were achieved for hiring decision and stress resistance, respectively. Results obtained from these two categories for the other hirability variables were however not much more accurate than the baseline-average model. Additionally, using the LIWC categories without word count and words per sentence are not predictive of any hirability score, as all R^2 values were negative using these predictors.

At first glance, these results suggest that the two linguistic categories of word count and words per sentence can be used for the prediction of hirability impressions, even if they were not as accurate as nonverbal behavior. However, word count and words per sentence cannot be considered as verbal features as such, because they do not encode the speech content, but the applicant's amount of speech and fluency. The real impact of verbal behavior in the formation of hirability impressions should be considered without these two (nonverbal) categories, and results showed that verbal behavior alone (LIWC without word count and words per sentence) is not predictive of hirability impressions.

Table 5.9 – Performance (R^2 and $RMSE$) for the inference of hirability scores using Linguistic Inquiry Word Count (LIWC) verbal categories combined with nonverbal behavior, using ridge regression and no dimensionality reduction. Acronyms: LIWC - all LIWC categories; LIWC (no WC/WPS) - all LIWC categories except word count and words per sentence; WC+WS - word count and words per sentence; NVB - interviewer and applicant nonverbal cues; APP - applicant nonverbal cues; APPAUDIO - applicant audio nonverbal cues. $N = 62$.

	HirDecision		**Communication**		**Persuasion**		**Conscience**		**StressRes**	
	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$
Baseline-Avg	0.00	1.79	0.00	0.98	0.00	1.03	0.00	0.95	0.00	0.79
LIWC	0.09	1.71	-0.05	1.00	-0.04	1.05	-0.05	0.97	0.05	0.77
LIWC (no WC/WPS)	-0.05	1.83	-0.05	1.00	-0.05	1.05	-0.05	0.97	-0.04	0.80
WC+WPS	0.28	1.52	0.06	0.95	0.02	1.02	0.08	0.91	0.16	0.72
NVB	0.36	1.43	-0.06	1.01	0.11	0.97	-0.05	0.97	0.13	0.74
NVB+LIWC	0.37	1.42	-0.06	1.00	0.12	0.96	0.03	0.93	0.12	0.74
NVB+LIWC (no WC/WPS)	0.34	1.46	-0.05	1.00	0.12	0.96	0.02	0.94	0.10	0.75
NVB+WC+WPS	0.40	1.39	-0.06	1.01	0.11	0.97	-0.03	0.96	0.14	0.73
APP	0.29	1.51	-0.05	1.00	0.06	0.99	-0.03	0.96	0.10	0.75
APP+LIWC	0.37	1.42	-0.02	0.99	0.08	0.99	-0.01	0.95	0.17	0.72
APP+LIWC (no WC/WPS)	0.32	1.48	-0.05	1.00	0.07	0.99	-0.06	0.97	0.15	0.73
APP+WC+WPS	0.36	1.43	-0.05	1.00	0.05	1.00	0.00	0.94	0.13	0.74
APPAUDIO	0.34	1.46	-0.04	0.99	0.08	0.99	-0.01	0.95	0.13	0.74
APPAUDIO+LIWC	0.40	1.39	-0.04	1.00	0.07	0.99	-0.02	0.96	0.18	0.72
APPAUDIO+LIWC (no WC/WPS)	0.32	1.47	-0.04	1.00	0.06	1.00	-0.02	0.96	0.14	0.73
APPAUDIO+WC+WPS	0.39	1.40	-0.04	1.00	0.07	0.99	0.00	0.94	0.15	0.73

5.6.3 Combining verbal and nonverbal cues

To understand whether verbal behavior can improve the inference of hirability impressions when combined with nonverbal behavior, we conducted a prediction experiment where LIWC features were fused with the nonverbal cues extracted in Section 4.2 (**NVB-CUES** feature subset). The fusion was achieved by concatenating the feature vectors. Specifically, we combined LIWC features with all (*i.e.* interviewer and applicant), applicant, and applicant audio cues. Inference was obtained using ridge regression. Results are shown in Table 5.9, where ALL, APP, and APPAUDIO denote interviewer and applicant cues, applicant cues, and applicant audio cues, respectively.

For the variable of hiring decision, the best prediction score was obtained for interviewer and applicant nonverbal cues combined with word count and words per sentence and applicant audio nonverbal cues with all LIWC categories ($R^2 = 0.40$). For stress resistance, applicant audio combined with all LIWC categories yielded the most accurate prediction result ($R^2 = 0.18$). For persuasion, the best accuracy was obtained for interviewer and applicant cues combined with LIWC categories including or excluding word count and words per sentence ($R^2 = 0.12$ for both). For the variables of communication and conscience, neither nonverbal nor verbal cues were able to yield prediction results more accurate than the baseline-average model.

For the hiring decision score, LIWC categories improved the prediction accuracy for applicant and interviewer nonverbal cues ($\Delta R^2 = 0.09$), applicant nonverbal cues ($\Delta R^2 = 0.08$), and applicant audio cues ($\Delta R^2 = 0.06$). Although these results suggest that verbal behavior might be useful to improve the prediction of hirability impressions, removing the nonverbal LIWC categories of word count and words per sentence resulted in an improvement only for

applicant nonverbal cues ($\Delta R^2 = 0.03$); for applicant and interviewer cues, and for applicant audio cues this resulted in a performance drop. Additionally, combining the two categories of word count and words per sentence to nonverbal cues improved the prediction performance ($\Delta R^2 = 0.03$ for all nonverbal cues, $\Delta R^2 = 0.07$ for applicant cues, and $\Delta R^2 = 0.05$ for applicant audio cues). This showcases the fact that in combination with nonverbal behavior, the only useful information contained in LIWC categories for the prediction of hirability impressions was contained in the two nonverbal categories of word count and words per sentence.

5.6.4 Discussion

In this section, we evaluated the use of verbal content represented by Linguistic Inquiry Word Count (LIWC) categories for the prediction of hirability impressions. To understand the linear relationship between LIWC categories and hirability, we first performed a correlation analysis. We observed that applicants who spoke more and used longer sentences were perceived as more hireable by raters. Applicants who used a large number of affective words (both positive and negative), as well as a simpler and more immediate speech (larger use of present tense, shorter words, and more elaborate words) received less positive hirability ratings.

For the task of automatically predicting hirability impressions, verbal content, represented by LIWC categories, was unable to yield results significantly higher than the baseline-average model. The two LIWC categories of word count and words per sentence taken alone shown better predictive validity, but still were less predictive than nonverbal behavior. When combined with nonverbal cues, LIWC features only marginally improved the prediction accuracy of hirability impressions, and removing word count and words per sentence resulted in a performance drop. These observations suggest that the relevant information for the task of hirability prediction was mainly contained in the two categories of word count and words per sentence. However, these two categories cannot be considered strictly as verbal cues, as they do not encode *what* was being said, but rather *how* the message was conveyed; in this sense, the only LIWC categories predictive of hirability impressions were nonverbal ones.

The two categories of word count and words per sentence are strongly related to the concept of fluency [38] (as a matter of fact, they were strongly correlated with applicant speaking time and average turn duration), and our prior results on nonverbal behavior analysis (Section 5.1), as well as previous studies in psychology have shown that nonverbal fluency was associated with hirability impressions [90, 45]. The fact that adding word count and words per sentence to nonverbal cues improved the prediction accuracy suggests that *verbal* fluency somewhat differs from *nonverbal* fluency; otherwise the features would have been redundant and would not have improved the prediction score. This finding suggests that word count and words per sentence could be seen as an alternative measure of applicant fluency.

A hypothesis to explain why verbal content as such was not predictive of hirability impressions could be that its representation (Linguistic Inquiry Word Count) was not the appropriate one for the task at hand. However, previous work on related constructs such as personality [31]

or leadership [121] have shown that LIWC could be used as a representation for automatic prediction. Investigating the use of alternative representations of verbal behavior, such as n-grams could be an interesting avenue for future work, but might be problematic due to the high dimensionality of the obtained feature vector in comparison to the number of subjects in this study. Another hypothesis to explain these results could be that raters made their impressions based on nonverbal behavior rather than verbal content. Previous work in personnel psychology [93] has suggested that nonverbal behavior can overshadow verbal first impressions in job interviews, providing evidence to this second hypothesis. A last hypothesis to explain the low predictive validity of verbal features could be that LIWC does not encode the linguistic alignment possibly present between the interviewer and the applicant. During dialogues, the words used by the protagonists tend to converge [107], and the amount of verbal synchrony could be predictive of hirability, but this hypothesis needs to be investigated in greater detail.

Due to the difficulties in finding a publicly available and accurate French automatic speech recognition (ASR) system, all analyses of verbal content were completed on labor-intensive and unscalable manual transcriptions. In a fully automated framework, one can question the effect of unavoidable ASR transcription errors on LIWC categories. In [31], authors showed that even the use of a state-of-the-art ASR can result in a dramatic performance drop in construct prediction using LIWC. Therefore, as in this study verbal behavior from perfect transcriptions was unable to accurately predict hirability impressions, we doubt that accurate prediction results can be achieved from error-prone automatic transcriptions. For this reason, we did not investigate the use of French ASR for this study.

5.7 Conclusion

In this chapter, we proposed a computational framework for the automatic prediction of hirability in real job interviews, using applicant and interviewer behavioral cues extracted from the audio and visual modalities. To our knowledge, this work is the first computational attempt aiming at systematically analyzing verbal and nonverbal behavior in job interviews, and the first focusing on the task of hirability prediction.

As a first step, we performed a correlation analysis between nonverbal cues and hirability impressions and found that not only applicant cues were correlated with the hirability scores, but interviewer cues, too. As a second step, we evaluated several prediction methods for a regression task. Results demonstrated the feasibility of predicting hirability scores based on automatically extracted nonverbal cues, and validated our proposed framework, with R^2 values of up to 36%. We then analyzed the predictive validity of feature groups and we observed that the most predictive groups were the applicant audio cues and the interviewer visual cues. This second finding suggests that the interviewer produced behavioral responses which were conditioned on the quality of the job applicant by displaying more visual back-channels. This observation shows the potential of predicting the interview outcome by looking

at the interviewer. We then analyzed the use of psychometric questionnaires widely used in the personnel selection process for the prediction of hirability scores. Questionnaires were unable to predict hirability more accurately than the baseline model. Moreover, combining the questionnaire scores to nonverbal cues did not improve the prediction accuracy compared to nonverbal behavior only.

We then conducted a systematic study on applicant body communication in job interviews with respect to hirability impressions and self-rated personality. We leveraged on findings in psychology suggesting a strong link between body communication and speech to analyze body communication from a multimodal perspective. By analyzing the corresponding differences in applicant body communication feature values, we validated our main assumption stating that speaking and silent differences existed. We showed that the prediction of interview outcomes using body communication cues was a promising task. We also showed that body communication cues can be used to predict applicant personality traits to a moderate degree, achieving results similar to the state of the art. The reported results also demonstrated that exploiting the intimate link between body communication and speaking status helped towards the inference of personality and hirability. The prediction of some of the constructs analyzed in this work relied on manual annotations of body activity. This was the case of personality traits, where no automatic feature could produce accurate prediction results. However, results showed that in certain cases, using the automatic hand speed estimates yielded higher prediction results than manual features. This was the case for the variable of hiring decision. This finding underlines the relevance of automatic hand speed estimates for the analysis of employment interviews, even if these estimates were coarse.

Last, we analyzed the predictive validity of verbal content. To this end, we relied on manual speech transcriptions and automatically extracted linguistic and paralinguistic features using the Linguistic Inquiry Word Count (LIWC) as a representation of verbal content. The correlation analysis showed that the two categories of word count and words per second were strongly associated with hirability impressions, confirming the finding that fluent applicants (*i.e.* applicants who spoke longer and had longer sentences) received high hirability ratings; these two categories however cannot be strictly considered as verbal because they do not encode the speech content, but rather the way speech is conveyed. Verbal content by itself was not able to yield accurate prediction results. When combined to nonverbal behavior, verbal categories were unable to improve the prediction accuracy. These results suggest that raters might have formed their hirability impressions based on nonverbal behavior rather than verbal content.

6 Thin slices of behavior in employment interviews

One of the objectives of this thesis is to understand how hirability impressions are formed during job interviews. To this end, the previous chapter investigated the use of various behavioral channels for the inference of hirability impressions based on the full interviews. In this chapter, we explore the amount of information which can be inferred from brief excerpts of behavioral stream, also known as *thin slices*. To this end, we used the SONVB interview dataset and utilized the structured nature of the interviews to segment each of them into slices of short durations. To understand the accuracy of raters when only a limited amount of information is available, we collected hirability annotations based on thin slices and correlated them with the ratings obtained from the full interview. Then, we extracted nonverbal features for both the applicant and the interviewer for each slice, and used them as predictors to infer full interview hirability impressions.

In our everyday lives, many decisions or judgments people make about others are made based on inferences arising from brief interactions. Social psychology research has shown that the proverb "first impressions are the ones lasting" holds true up to a certain extent: humans are quite accurate at making inferences about others, even if the information is minimal [19]. Short segments of interactions, typically under five minutes, are commonly referred to as thin slices [19]. Surprisingly, such minimal displays of behavior can be predictive of social constructs (*e.g.* personality, competence) and outcomes (*e.g.* teacher ratings) [19]. As an extreme example of thin-slicing, inferences of competence by naïve raters based on simple photographs were strongly correlated with election outcomes [132]. In job interviews, because protagonists are most often strangers, recruiters have access to very little information (usually, the verbal and nonverbal elements of the interaction, as well as the resume); in this type of setting, nonverbal behavior is known to play a key role [78], which makes job interviews an interesting case to examine the interplay between nonverbal behavior and thin slices.

Thin slice research in social psychology has examined the amount of information that could be inferred from short excerpts of behavior by unacquainted judges. To this end, the concept of *observer predictive validity* has been used as an assessment of the relationship between thin slice ratings and the ground truth, defined by direct measurements, self-reports, or

impressions obtained from the full interaction, depending on the abstraction level of the social construct being judged [17]. Although the most widely used measure for predictive validity is the correlation between thin slice ratings and the ground truth, other metrics exist, such as the amount of agreement among raters. Works in social psychology have shown that thin slices could be predictive of a broad array of social outcomes, such as individual performance (teaching, job performance, health care), relationships (type and quality of relationships), and individual differences (personality, gender, sexual orientation) [17]. Social psychology research has also investigated the predictive validity of thin slices depending on the channels of communication; the nonverbal channel was shown to play an important role in the formation of these first impressions in such brief excerpts of social interactions [19, 17].

Surprisingly, only a few studies have investigated the effect of impressions from thin slices on job interview outcomes. In an unpublished Master's thesis [55], unacquainted judges rated short pre-interview thin slices, defined by the 10 seconds following the moment when the job applicant took his seat; thin slices ratings were observed to be significantly correlated with the full interview ratings. Another work [123] examined the relationships of hiring recommendations based on thin slices and full interviews, and ratings based on 12-second silenced snippets of video were correlated with full interview ratings. Impressions of social skills (*e.g.* attentive, anxious, confident, etc.) based on thin slices were observed to be associated with full interview ratings, whereas manually annotated visual nonverbal behaviors (*e.g.* head nods, smiles, fidgets, etc.) did not show any significant correlation.

Several computational studies have examined the use of thin slices in contexts similar to job interviews. The work in [44] studied the relationship between nonverbal behavior and outcomes in a simulated dyadic negotiation scenario, focusing on the first five minutes of the interaction. In [25], personality traits were predicted from self-presentations ranging from 30 to 120 seconds in a human-computer setting. Other computational studies have investigated the use of thin slices for the prediction of social constructs as diverse as interest [86], personality [27, 106], attraction [86], emergent leadership [119], or individual performance [81] in face-to-face interactions. In most of these works [86, 25, 81], the concept of thin slices was used because the interactions were inherently short, and both the extracted behavioral features and the annotations of social variables stemmed from the full interactions. However, the study in [119] investigated the effect of slice durations on the prediction accuracy by extracting behavioral features from the slice of interest and using them to infer the social variables annotated from the full interaction, and observed that an asymptote was reached around the middle of the interaction.

To the best of our knowledge, no computational study has examined the role of thin slices in employment interviews. We believe that job interviews are an interesting setting to investigate the interplay between thin slices and nonverbal behavior. This chapter specifically aims to address the following research questions: (Q1) Can a short excerpt of the interview be predictive of the hirability ratings based on the full interview? (Q2) If so, what are the most predictive slices? (Q3) What are the cues used for prediction, and are they consistent across

Slices were defined based on the questions of the structured interview. The sequence of questions is listed below:

1. Short self-presentation.
2. Motivation for applying for the job.
3. Importance of scientific research (which is the field of the job).
4. Past experience where communication skills were required.
5. Past experience where persuasion skills were required.
6. Past experience of conscientious/serious work.
7. Past experience where stress was correctly managed.
8. Strong/weak points about self.

Figure 6.1 – Interview structure.

slices? (Q4) Is the interaction during interview questions predictive of hirability compared to the interview answers? We approach these research questions from a nonverbal perspective where sensing, cue extraction, and prediction are automated. The study has the potential to improve the understanding of the predictive validity of questions usually posed during employment interviews.

The material in this chapter has not been published elsewhere.

6.1 Data and annotations

6.1.1 Data and full interview annotations

To address our research questions, we used the 62 job interviews of the SONVB dataset (**SONVB-ALL** data subset), described in Section 3. Hirability impressions were collected by a pool of five human resource professionals who watched the full interviews including the audio track; all videos were annotated by three raters (**HR** annotation scheme - see Section 3.4). Only the hiring decision variable was used in this chapter because we believe that it constitutes the most relevant hirability variable. Analyses were also completed for the other hirability scores but did not drastically differ with respect to the hiring decision variable, so they were not included here.

As a reminder, job interviews were structured, meaning that they followed the same sequence of questions and answers to ensure that comparisons could be made between participants. In total, the job interviews consisted of eight questions, which are listed in Figure 6.1. More details about the scenario can be found in Section 3.

6.1.2 Definition of thin slices

Most previous studies in thin slices used segments of fixed duration, which is valid for the case of unstructured interactions, but generates an undesirable bias when the interaction is structured such as in our case. To prevent this, we decided to make use of the structured nature of the interviews by annotating the timings of the eight specific question/answer segments of the job interview. Additionally, in order to compare the behavioral predictive power of

Table 6.1 – Definition of the three thin slice cases used for the eight slices of the structured interview.

TS case	$\mathbf{t_{start}}$	$\mathbf{t_{end}}$
Question-only	Question start	Question end
Answer-only	Answer start	Answer end
Whole-slice	Question start	Answer end

Table 6.2 – Statistics on the duration of slices: mean and standard deviation (in seconds). $N = 62$.

		Question-only		**Answer-only**		**Whole-slice**	
Slice	Slice description	mean [s]	std [s]	mean [s]	std [s]	mean [s]	std [s]
1	Self-presentation	64.5	6.1	63.4	34.8	128.0	36.2
2	Motivation for the job	3.3	0.5	35.3	18.4	38.6	18.3
3	Importance of research	5.4	0.7	46.4	20.8	51.8	20.8
4	Communication skills	31.2	2.3	64.2	33.3	95.4	33.7
5	Persuasion skills	20.7	1.5	60.7	32.9	81.5	33.2
6	Conscientiousness	27.8	1.9	55.4	34.6	83.3	34.7
7	Stress resistance	15.4	1.1	60.8	43.2	76.2	43.0
8	Strong/weak points	6.0	0.7	71.6	32.5	77.6	32.3

question *vs.* answer segments, we further annotated the timing of questions and answers. The annotations of timings were completed by the PhD candidate. In summary, three thin-slice cases were used: whole-slice, question-only, and answer-only (see Table 6.1).

Statistics of slice durations are shown in Table 6.2. We first observe that the longest slice (whole-slice) is the first question (self-presentation), which can be explained by the fact that the question was significantly longer than the other ones because it included the job description. The low variance in duration for the question-only case can be explained by the fact that the interviewer was instructed to follow a script. In terms of answers, six had an average duration over 50 seconds. The shortest answer was the motivation to apply for the job, and this can be explained by the fact that some applicants had already mentioned their motivation in the first question (self-presentation).

6.1.3 Hirability impressions from thin slices

To assess the accuracy of human raters when exposed to short excerpts of interactions *vs.* when they have access to the full interviews, we collected annotations for each interview slice. To this end, a pool of four human resource professionals rated each individual slice (whole-slice - including question and answer), where audio was also included. In total, each rater watched two slices from the same interview; the two slices were ensured not to be subsequent such that the hirability impression on the second slice was not too heavily influenced by the first one. Raters were instructed to give their hiring decision impression on the interview thin slice, which corresponds to the thin-slice equivalent of the hiring decision score introduced in

Section 3.4.

To assess agreement between judges, all 8 slices were viewed by a second rater for 10 interviews. Inter-rater agreement was computed for each slice using Pearson's pairwise correlation coefficient, and results are displayed in Figure 6.2(1); the inter-rater agreement for the full interview (using the average Pearson's correlation coefficient as measure) is also displayed. Although the number of double-coded videos for each slice was relatively low, it provides a good insight on the ability of judges to form impressions from short excerpts of interviews. Inter-rater agreement for slice 4 (communication skills) and slice 8 (strong/weak points) was observed to be low, which suggests that these slices were more difficult to annotate. Other slices had agreements ranging from $r = 0.50$ to $r = 0.85$.

As a second step, we computed the pairwise correlation between slice and full interview annotations, using Pearson's correlation coefficient. Slice-full correlations are displayed in Figure 6.2(2). In social psychology related works [17], this measure is often used to quantify the observer predictive validity of the slice, which corresponds to how much an unacquainted judge can infer from a short excerpt of behavioral stream. Slice-full correlations ranged from $r = 0.25$ to $r = 0.69$. We observe that not all slices showed the same observer predictive validity: slice 2 (motivation for the job) and slice 4 (communication skills) were found to be less strongly correlated with the full interview rating ($r = 0.36$ and $r = 0.25$, resp.) than other slices. The low predictive validity of slice 4 (communication skills) can be explained by the fact that the agreement among judges for this slice was low ($r = 0.05$). For slice 2 (motivation for the job), although the inter-rater agreement was high ($r = 0.84$) the slice showed poor observer predictive validity. This finding can be explained by the fact that some job applicants stated their motivation to apply for the job in the previous question (self-presentation) and provided a short answer (*e.g.*, "As I already told you...") for slice 2, possibly earning low hirability ratings due to this.

As a last step, we computed the squared value of the slice-full correlations. The obtained R^2 value accounts for the variance explained by an ordinary least squares linear model using the slice annotation as only predictor, with no cross-validation. In other words, this number represents the amount of explained variance by holding the rating based on the thin slice, and could be seen as an upper bound for the automatic prediction of full-interview hirability impressions based on thin slices. The R^2 values for each slice are displayed in Figure 6.2(3). We observe that the obtained R^2 values range from $R^2 = 0.06$ (communication skills) to $R^2 = 0.47$ (stress resistance). As a comparison, experiments competed in Section 5.3 using nonverbal features yielded accuracies up to $R^2 = 0.40$, suggesting that using an automated method might produce results similar to human observations of thin slices.

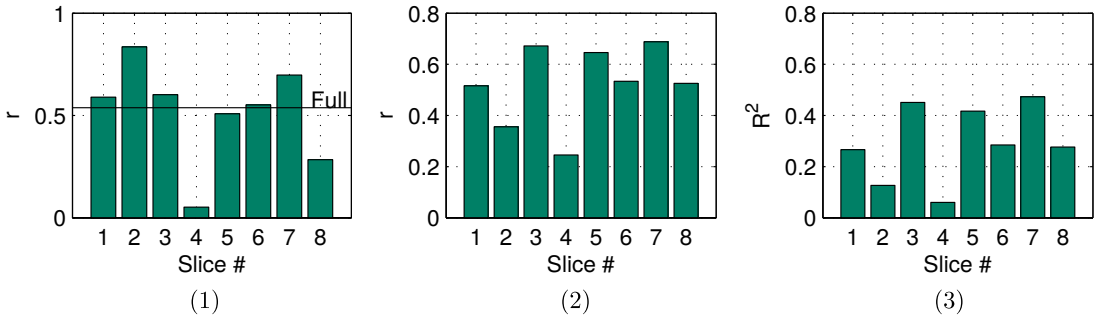


Figure 6.2 – (1) Slice-level inter-rater agreement ($N = 10$), using Pearson's correlation as agreement measure. The solid line denotes the average correlation between the three raters on the full interview ($N = 62$). (2) Pearson's correlation between thin slice and full-interview annotations ($N = 62$). (3) R^2 values for each slice (squared of values displayed in (2)), corresponding to the prediction accuracy using only the slice annotations and OLS regression model ($N = 62$).

6.2 Behavioral features

6.2.1 Extracted features

To understand whether hirability impressions based on full interviews could be inferred from thin slices, we extracted behavioral cues for each slice of the interview. Because they were shown to be useful for the prediction of hirability impressions from full interviews, we used the audio-visual nonverbal cues presented in Section 4.2 (**NVB-CUES** feature subset). In this chapter, we re-extracted these nonverbal cues for the eight interview questions and the three thin slice cases (see Table 6.1), for a total of 24 slice-cases. To attenuate the effect of variable slice lengths, features were normalized with respect to the duration of the slice.

6.2.2 Correlation analysis

Pairwise correlations between single behavioral cues extracted from thin slices and hirability impressions obtained from the full interview were computed, and are displayed in Tables 6.3 and 6.4. We now discuss the main findings.

Applicant cues

In Table 6.3, we display the applicant cues significantly correlated with the full interview hiring decision score, where cues were extracted from thin slices (whole-slice). For comparison, nonverbal cues extracted from the full interview are also displayed.

Some applicant audio features were found to be consistently and significantly correlated with the full interview hiring decision across slices. This is the case of the prosodic cues related to energy (median and lower quartile), voiced rate (mean, std, median, and upper quartile), and pitch (std, median, and lower quartile), which were found to be associated with the hiring

decision score, when extracted from both the full interview and most thin slices.

For applicant cues based on speaker segmentations, the number of turns, silence features (although also related to the interviewer), and short utterances were observed to be negatively correlated with the hiring decision score across slices and for the full interview case; applicant average turn duration was positively associated with the interview outcome for the full interview and some slices. This observation suggests that these behavioral cues were consistently displayed by job applicants. This was however not the case of applicant speaking time, which was positively associated with the hiring decision for the full interview case, but negatively correlated for most thin slices. This is an interesting result because speaking time has been shown to be a robust predictor of other social constructs including personality [78].

To a lesser degree, applicant vertical head motion (mean and median) was positively associated with the hiring decision score, independently from the fact that they were extracted from the full interview or the thin slices. Otherwise, no applicant visual behavioral cue was consistently correlated with the interview outcome.

In terms of the number of significantly correlated features with the hiring decision score, all slices were not equal. Slice 4 (communication skills) was the slice from which the larger number of applicant cues significantly correlated with the hiring decision were extracted (32 significantly correlated cues whereas other slices range from 18 to 25). This finding is paradoxical because this slice was the one showing the lowest observer predictive validity ($r = 0.25$), and was also the slice where the agreement among raters was the lowest ($r = 0.05$).

Interviewer cues

In Table 6.4, we display the interviewer cues that were significantly correlated with the full interview hiring decision score, where cues were extracted from thin slices (whole-slice case). For comparison, the correlations for the nonverbal cues extracted from the full interview are also displayed.

Interviewer pitch standard variation was observed to be consistently and negatively associated with the hiring decision score across slices. This suggests that the interviewer had a more monotonous tone of voice in presence of highly hirable job applicants, and that this behavior was displayed throughout the totality of the job interview, with the exception of slice 1 (self-presentation).

Interviewer short utterances were also observed to be negatively and consistently correlated with the interview hiring decision score. This finding corroborates the one in Section 5.1: the short utterances stemmed from short back-and-forth exchanges between the applicant and the interviewer and could be seen as clarifications asked by the job applicant. These short utterances were perceived negatively by raters, and the effect can be observed throughout the full interview.

Table 6.3 – Applicant nonverbal cues extracted from thin slices (whole-slices) and from the full interview significantly correlated with full interview hiring decision score ($p < .05$, $\dagger p < .005$). $N = 62$.

Feature	Full	Slice 1	Slice 2	Slice 3	Slice 4	Slice 5	Slice 6	Slice 7	Slice 8
Applicant prosodic features (audio):									
Applicant pitch std	$-0.44^{\dagger}$	$-0.39^{\dagger}$	−0.29	$-0.42^{\dagger}$	$-0.48^{\dagger}$	$-0.44^{\dagger}$	$-0.45^{\dagger}$	$-0.39^{\dagger}$	$-0.41^{\dagger}$
Applicant pitch median	0.28	0.25		0.28	0.28	0.26	0.29	0.28	0.31
Applicant pitch lower quartile	$0.39^{\dagger}$	0.35	$0.36^{\dagger}$	$0.41^{\dagger}$	$0.40^{\dagger}$	$0.40^{\dagger}$	$0.39^{\dagger}$	$0.37^{\dagger}$	$0.37^{\dagger}$
Applicant energy median	0.33	0.31	0.23	0.29	0.34	0.34	0.32	0.28	
Applicant energy lower quartile	$0.37^{\dagger}$	$0.43^{\dagger}$	$0.36^{\dagger}$	0.35	0.35	0.34	0.33	0.35	
Applicant energy upper quartile				0.27					
Applicant voiced rate mean	$0.54^{\dagger}$	$0.38^{\dagger}$	$0.42^{\dagger}$		$0.36^{\dagger}$	$0.44^{\dagger}$	$0.47^{\dagger}$	$0.53^{\dagger}$	0.29
Applicant voiced rate std	$0.50^{\dagger}$		0.30		0.30		$0.42^{\dagger}$	$0.44^{\dagger}$	
Applicant voiced rate median	0.32		$0.43^{\dagger}$	0.32		0.35	0.34	$0.40^{\dagger}$	
Applicant voiced rate lower quartile					0.28				
Applicant voiced rate upper quartile	$0.47^{\dagger}$	0.29	$0.40^{\dagger}$		$0.37^{\dagger}$	0.29	$0.41^{\dagger}$	$0.42^{\dagger}$	0.29
Applicant voiced rate maximum			0.28		0.32		0.34	0.35	
Applicant turn based features (audio):									
Applicant # of turns	$-0.58^{\dagger}$	$-0.39^{\dagger}$	$-0.38^{\dagger}$	$-0.37^{\dagger}$	$-0.38^{\dagger}$	$-0.37^{\dagger}$	−0.33	$-0.38^{\dagger}$	$-0.38^{\dagger}$
Applicant speaking time	$0.48^{\dagger}$		−0.33		−0.31	−0.30		−0.32	−0.30
Applicant average turn duration	$0.45^{\dagger}$	$0.39^{\dagger}$		$0.50^{\dagger}$		0.29			
Applicant maximum turn duration	$0.53^{\dagger}$		−0.31						−0.27
Applicant number of pauses		−0.26	−0.26		$-0.44^{\dagger}$	−0.35	−0.33	−0.33	−0.26
Number of silent segments	$-0.50^{\dagger}$	$-0.39^{\dagger}$	$-0.43^{\dagger}$	$-0.39^{\dagger}$	$-0.46^{\dagger}$	$-0.37^{\dagger}$	$-0.46^{\dagger}$	$-0.40^{\dagger}$	$-0.44^{\dagger}$
Total silent time	$-0.58^{\dagger}$	$-0.41^{\dagger}$	$-0.41^{\dagger}$	$-0.37^{\dagger}$	$-0.48^{\dagger}$	$-0.38^{\dagger}$	$-0.42^{\dagger}$	$-0.43^{\dagger}$	$-0.46^{\dagger}$
Number of overlapping segments		−0.32		−0.31	−0.29			−0.31	$-0.36^{\dagger}$
Total overlapping time		−0.33		−0.30	−0.32			−0.35	$-0.37^{\dagger}$
Applicant number of short utterances	$-0.45^{\dagger}$	−0.35	$-0.43^{\dagger}$	$-0.38^{\dagger}$	$-0.43^{\dagger}$	$-0.37^{\dagger}$		$-0.37^{\dagger}$	$-0.45^{\dagger}$
Applicant total short utterance time	$-0.45^{\dagger}$	−0.35	$-0.46^{\dagger}$	$-0.39^{\dagger}$	$-0.43^{\dagger}$	$-0.37^{\dagger}$		$-0.38^{\dagger}$	$-0.44^{\dagger}$
Applicant # of audio back-channels		−0.32			−0.27				−0.32
Applicant audio back-channel time		−0.32			−0.28				$-0.36^{\dagger}$
Applicant WMEI features (visual):									
Applicant WMEI vertical center of mass							0.26		
Applicant WMEI non-zero ratio	−0.29				−0.26				−0.34
Applicant nod-based features (visual):									
Applicant number of head nods		−0.28		−0.28	−0.32	−0.26		−0.30	−0.30
Applicant nodding time					−0.28	−0.28		−0.27	
Applicant number of visual back-channel events		−0.29			−0.30	−0.28		−0.27	−0.33
Applicant visual back-channeling time	0.26					−0.29		−0.27	−0.31
Applicant number of nodding while speaking events				−0.26	−0.29				
Applicant head motion features (visual):									
Applicant median horizontal head motion		0.29							
Applicant maximum horizontal head motion				−0.31	−0.26				
Applicant mean vertical motion	0.31	0.31	0.26		0.29	0.31	$0.37^{\dagger}$		
Applicant median vertical head motion	$0.40^{\dagger}$	$0.40^{\dagger}$		0.33	$0.37^{\dagger}$	0.35	$0.44^{\dagger}$		
Applicant maximum vertical head motion	−0.25				−0.30				
Applicant mean head motion magnitude		0.29							
Applicant median head motion magnitude	0.31	$0.38^{\dagger}$		0.26	0.25		0.31		
Applicant maximum head motion magnitude			0.26	−0.28	−0.30				

Table 6.4 – Interviewer nonverbal cues extracted from thin slices (whole-slices) and from the full interview significantly correlated with full interview hiring decision score ($p < .05$, $\dagger p < .005$). $N = 62$.

Feature	Full	Slice 1	Slice 2	Slice 3	Slice 4	Slice 5	Slice 6	Slice 7	Slice 8
Interviewer prosodic features (audio):									
Interviewer pitch std	$-0.42^{\dagger}$		-0.33	$-0.36^{\dagger}$	-0.30	-0.26	-0.32	-0.32	-0.28
Interviewer pitch lower quartile				0.34					0.32
Interviewer pitch upper quartile			-0.30						
Interviewer pitch maximum				-0.26					
Interviewer energy median				0.25					
Interviewer energy lower quartile	0.30			$0.37^{\dagger}$					
Interviewer voiced rate mean	0.27								
Interviewer voiced rate median	0.30								
Interviewer voiced rate lower quartile							0.27		
Interviewer turn based features (audio):									
Interviewer number of turns		-0.34	-0.30	-0.34	-0.33	-0.30	-0.30	-0.34	-0.30
Interviewer speaking time		-0.33	-0.31	-0.31	-0.33	-0.33	-0.32	-0.33	-0.34
Interviewer average turn duration									-0.31
Interviewer maximum turn duration		-0.32	-0.31	-0.28	-0.32	-0.32	-0.31	-0.32	-0.33
Interviewer number of pauses		$-0.37^{\dagger}$			-0.29				
Number of silent segments	$-0.50^{\dagger}$	$-0.39^{\dagger}$	$-0.43^{\dagger}$	$-0.39^{\dagger}$	$-0.46^{\dagger}$	$-0.37^{\dagger}$	$-0.46^{\dagger}$	$-0.40^{\dagger}$	$-0.44^{\dagger}$
Total silent time	$-0.58^{\dagger}$	$-0.41^{\dagger}$	$-0.41^{\dagger}$	$-0.37^{\dagger}$	$-0.48^{\dagger}$	$-0.38^{\dagger}$	$-0.42^{\dagger}$	$-0.43^{\dagger}$	$-0.46^{\dagger}$
Number of overlapping segments		-0.32		-0.31	-0.29			-0.31	$-0.36^{\dagger}$
Total overlapping time		-0.33		-0.30	-0.32			-0.35	$-0.37^{\dagger}$
Interviewer number of short utterances	-0.33	-0.33	-0.33	$-0.36^{\dagger}$	$-0.39^{\dagger}$		-0.31		$-0.42^{\dagger}$
Interviewer short utterance time	$-0.37^{\dagger}$	-0.33	-0.34	$-0.37^{\dagger}$	$-0.39^{\dagger}$		-0.32		$-0.41^{\dagger}$
Interviewer number of audio back-channel events		-0.33		-0.30	-0.32		-0.28		
Interviewer audio back-channeling time		-0.33		-0.30	-0.33		-0.30		
Interviewer WMEI features (visual):									
Interviewer WMEI upper quartile					-0.26				
Interviewer nod-based features (visual):									
Interviewer number of head nods	0.35		-0.32	-0.29	-0.29	-0.29	-0.27	-0.31	-0.31
Interviewer nodding time	$0.42^{\dagger}$		-0.32		-0.29	-0.29		-0.29	-0.30
Interviewer number of visual back-channel events	$0.51^{\dagger}$		-0.32		-0.27	-0.26	-0.26	-0.29	-0.28
Interviewer visual back-channeling time	$0.54^{\dagger}$		-0.31		-0.28	-0.26		-0.27	-0.27
Interviewer number of nodding while speaking events		-0.35				-0.32			
Interviewer nodding while speaking time		-0.33				-0.32			
Interviewer head motion features (visual):									
Interviewer mean horizontal head motion			0.27	0.26					
Interviewer median horizontal head motion			0.27						
Interviewer mean vertical head motion	0.26			0.35					
Interviewer median vertical head motion	0.31			$0.37^{\dagger}$		0.31			
Interviewer mean head motion magnitude				0.33					
Interviewer median head motion magnitude	0.28		0.26	0.33		0.28			

Interestingly, head-nod-related cues (number of nods, nodding time, visual back-channels) were positively and significantly correlated to the hiring decision score when extracted from the full interaction; however, this tendency was reversed when the cues were extracted from thin slices. One possible hypothesis to explain this finding could be that these features were unstable when extracted over short periods. This issue of temporal stability needs to be examined in more detail.

6.3 Inference

6.3.1 Experiments

We defined the prediction task as a regression task, where the goal was to infer the hirability scores obtained from the full interview (**HR** annotation scheme - see Section 3.4) from behavioral features extracted from thin slices. As a possible prediction task, inferring the hirability impressions obtained from thin slices was also considered, but we decided not to address this task as we strongly believe that inferring slice impressions is not as useful as predicting the full interview outcome, even if they are correlated.

For prediction, we used ridge regression with no dimensionality reduction, as it was observed in Section 5.2 to be the most accurate method for the inference of hirability impressions for the full interview case. Other regression methods were also tested (random forest, LASSO regression, ordinary least squares) but yielded inferior prediction accuracies so they are not presented here. As before, we used leave-one-interview-out cross-validation, using all interviews except one for training, and the remaining one to evaluate our method. Prior to the inference step, all nonverbal features were normalized using the z-score; also, highly skewed features ($skewness > 1$) were transformed using log-transformation.

We ran the prediction task for all eight interview slices and for all three thin slice case (see Table 6.1), for a total of 24 slice-cases. Furthermore, to investigate what group of cues were predictive of hirability, we did the same experiment using feature groups based on modality and person of interest. Modality-based feature groups included 'audio', 'video', and 'all' (*i.e.* the combination of audio, video, and multimodal features from Section 4.2). Person-based feature groups included 'applicant', 'interviewer', and 'dyad' (*i.e.* the combination of applicant and interviewer features). As an evaluation measure, we used the coefficient of determination R^2.

6.3.2 Results and discussion

The prediction results obtained for each thin slice, person-based feature groups, modality-based feature groups, and thin slice cases, as well as the results obtained using the full interview are shown in Figure 6.3. The results show that all slices could be predictive of hirability ratings up to a certain level. This finding provides an answer to Q1, our first research question: a short

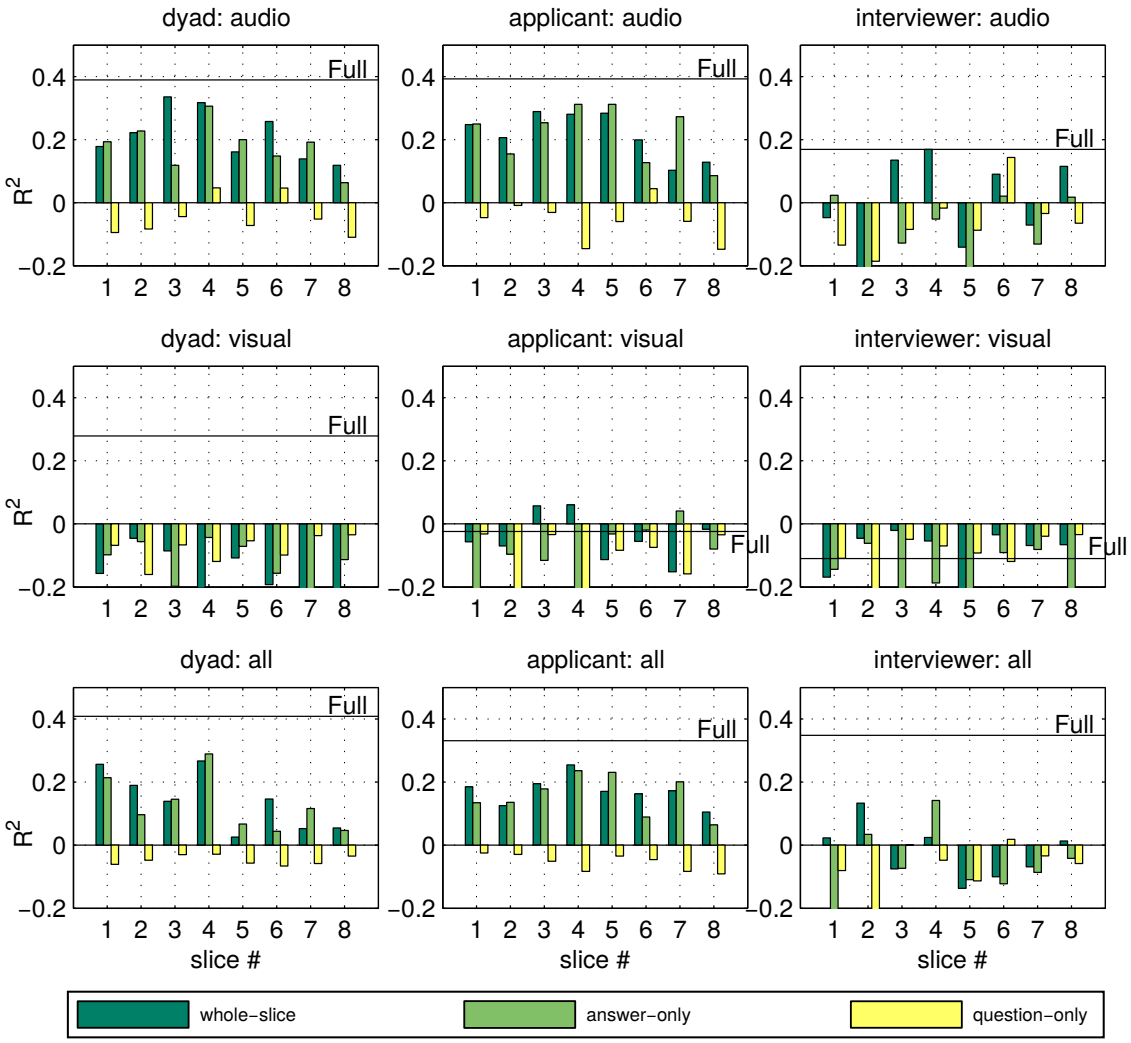


Figure 6.3 – R^2 results from thin slices for person-based feature groups (columns), modality-based feature groups (rows), and thin slice cases (dark green for whole-slice, light green for answer-only, and yellow for question-only). The solid line refers to the prediction result obtained using the full interaction. $N = 62$.

excerpt of a job interview can be predictive of hirability. The best results obtained from thin slices were competitive compared to the observer predictive validity, with R^2 of up to 0.34. However, in most cases the results obtained from thin slices were less accurate than the ones yielding from the full interview, suggesting that a larger amount of behavioral information remains beneficial for a better prediction.

For feature groups yielding positive prediction results for thin slices, no slice clearly stood out either negatively or positively. For the 'dyad:all' feature group, the second half of the interview (slices 5-8) tended to be slightly less predictive than the first half, but this was not observed for the other feature groups. Hence, no firm conclusion can be drawn on which question of the job interview elicited the most discriminative behavior for the prediction of hirability. These findings answer Q2, our second research question: no slice was clearly more predictive than the others.

We observe that the predictive validity of thin slices was dependent on both the modality and the person of interest. Applicant and dyad audio features extracted for the full interview were predictive of hirability ratings ($R^2 = 0.39$ for both). For thin slices, their prediction accuracy was consistent across segments of the interview. Interviewer audio cues were somewhat predictive for the full interview ($R^2 = 0.17$), but the validity of thin slices was not observed, as only slices 3, 4, 6, and 8 yielded mildly positive results. Visual features taken for the full interview were predictive of hirability for the dyad case ($R^2 = 0.28$), but were not predictive when thin-sliced, suggesting that these cues require longer temporal support to be predictive of the interview outcome. Applicant and interviewer visual features taken separately were not predictive of hirability for both the full interview and the thin slices. For head nods taken separately (not shown in Figure 6.3), interviewer and dyad nods were predictive when extracted from the full interview ($R^2 = 0.43$ and 0.40, resp.), but prediction accuracy dropped below zero for thin slices. This finding can be explained by the results obtained in Section 6.2.2, where interviewer heads nods extracted from the full interview were observed to be positively correlated with the hiring decision, while the ones obtained from thin slices were negatively correlated. This inconsistency in the display of head nods by the interviewer was responsible for the performance drop observed when using thin slices. These observations answer Q3, our third research question: only applicant and dyad audio cues were consistently predictive of hirability across slices.

In terms of thin slice cases, question-only segments were consistently not predictive of hirability, with frequent negative R^2 and quasi-constant poorer results compared to answer-only or whole-slice segments. The short duration of questions (see Table 6.2) fails to fully explain this finding, as the longest question (slice 1) was in no way more predictive than shorter answers. Along the same lines, whole-slices were not more predictive than answer-only segments, underlining the finding that questions did not add behavioral information to answers. These findings answer Q4, our fourth research question investigating the differences between questions and answers in terms of behavioral predictive validity: hirability ratings were not influenced by the listening behavior of applicants, but stemmed almost entirely from answers.

6.4 Conclusion

In this chapter, we analyzed thin slices of job interviews, where slices were defined by the specific questions of the interview structure. We used the full SONVB interview dataset comprising 62 job interviews. To assess the observer predictive validity of thin slices, annotations were completed on snippets of videos defined by the interview question/answer slices, and the slice annotations were analyzed using two measures, namely inter-rater agreement and slice-full correlations. Inter-rater agreement was observed to be greater than $r = 0.5$ for six out of eight slices, suggesting that most slices were rated consistently across raters. Six of eight slices also showed slice-full correlations over $r = 0.5$. Interestingly, one of the slices with low slice-full correlation was found to have high inter-rater agreement, which suggests that high agreement among judges is necessary, but not sufficient for satisfactory observer predictive validity on thin slices. Slice-full correlation coefficients values were then squared, yielding the amount of variance explained using an ordinary least squares linear model with the slice annotation as only predictor. The results showed that predicting hirability from automatically extracted nonverbal cues from the full interview yielded results similar to using annotations obtained by human resource professionals based on thin slices.

We then extracted nonverbal behavioral cues from the audio and visual modalities for both the applicant and the interviewer; these cues were extracted for all slices, as well as for the full interview. Applicant cues related to voice characteristics, speaking turns, and head motion were found to be significantly and consistently correlated with the hiring decision, with the exception of applicant speaking time which was positively correlated for the full interview, but negatively correlated for some slices. Some interviewer cues were also found to be consistently correlated with the full interview hiring decision score; this is the case of pitch standard deviation and short utterances. The interviewer's nodding behavior was observed to be conflicting: interviewer nods taken from the full interaction were positively correlated, while the valence was reversed in thin slices. The issue of temporal stability of some of these features need to be investigated in more detail.

We then performed a regression task, where the goal was to infer the full interview ratings based on the nonverbal cues extracted from the full slices. We used ridge regression and leave-one-interview-out cross-validation strategy to analyze the behavior predictive validity of thin slices, and although behavioral cues extracted from thin slices were found not to be as accurate as the full interaction, they were still predictive of the interview outcome: the best results obtained from thin slices were competitive compared to the observer predictive validity, with R^2 of up to 0.34. Furthermore, no slice stood out in terms of predictive validity: all slices yielded comparable results, suggesting that the observed nonverbal behavior did not drastically change from one slice to another.

To understand the basis on which raters formed their hirability impressions, we examined the accuracy stemming from person- and modality-based feature groups, and applicant audio features were observed to yield the most accurate results. To examine the predictive validity of

interview questions, we split the slices into three thin slice cases, question-only, answer-only, and whole-slice. Questions taken alone were found to consistently yield negative results, whereas answers predicted the hiring decision score to a lesser degree than using the full interview; moreover, adding the questions to the answers (*i.e.* using the whole-slice case) did not significantly improve the predictive validity, which suggests that raters made their impression based on the applicant's speaking behavior.

7 Hirability in the wild: analysis of online conversational video resumes

Online social media is changing the landscape of personnel recruitment. Beyond the massive success of LinkedIn and its 300+ million users from 200 countries [126], video interviews are modifying the way applicants get hired. In the last decade, numerous online video interviewing platforms such as HireVue [5], SparkHire [10], HireIQ [4], or VideoRecruit [11] have emerged and recruiters now not only can interview applicants online, but their presence is not even required anymore: applicants answer a set of predefined questions and their responses are audio- and video-recorded. In 2012, over 60% of the U.S. companies conducted video job interviews, and this number is expected to grow in the next years [33].

Until now, resumes were among the most widely used tools for screening job applicants in the personnel selection process [109]. The advent of inexpensive sensors (webcams, microphones) combined with the success of online video hosting and viewing platforms (*e.g.*, YouTube) has enabled the introduction of a new type of resume, the video resume. Video resumes can be defined as short video messages in which job applicants present themselves to potential employers [65]. Compared to traditional paper resumes, video resumes have the main advantage of providing the applicants with the opportunity to show their potential to the employer by displaying behavioral information through the audio and video channels [65]. The first video resumes appeared in the 1980s and were recorded on and distributed by VHS cassettes [42], but due to the lack of resources in production and distribution, the trend did not take off. In the last decade, conversational video resumes began to become popular among high school students who started sending their video resumes to colleges, in addition to their paper resume [42]. Presently, video resumes constitute an emerging phenomenon.

In this chapter, we analyze the formation of job-related first impressions in online conversational video resumes. This study constitutes the first opportunity to study organizational social constructs at a scale never previously achieved. We approach this problem from a nonverbal perspective, where feature extraction and statistical inference are fully automated. We collected a dataset of 939 conversational English-speaking video resumes from YouTube. Annotations of demographics, skills, and first impressions were collected using the Amazon Mechanical Turk (MTurk) crowdsourcing platform [1]. Basic demographics were then analyzed

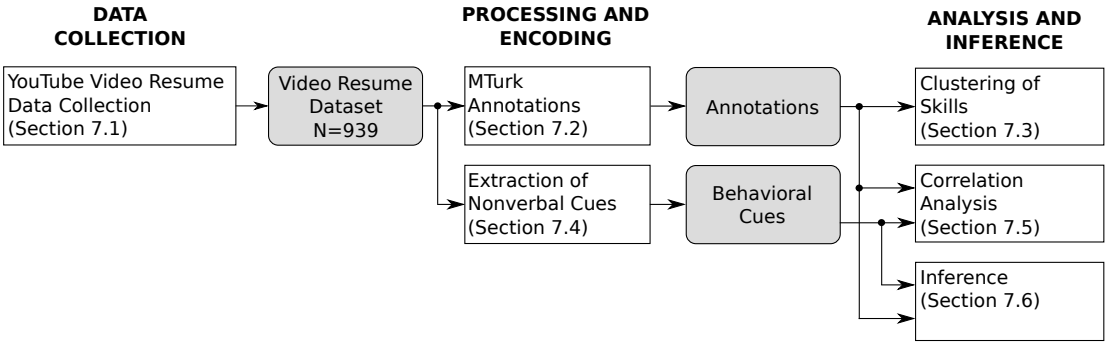


Figure 7.1 – Summary of Chapter 7.

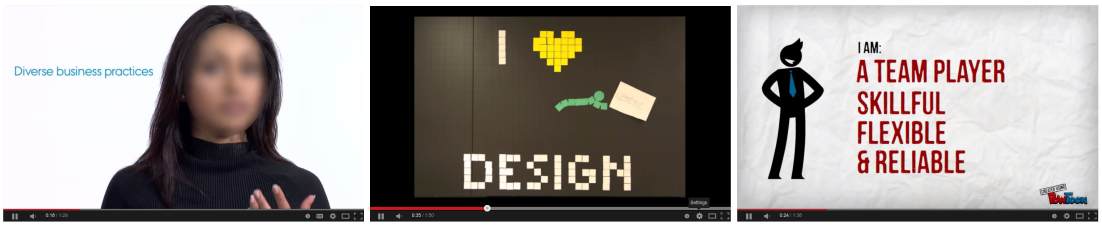
to understand the population using video resumes to find a job. The structure of perceived job-related skills was then analyzed in a data-driven clustering experiment. To obtain a feature representation of the video resumes, we extracted nonverbal cues from the audio and visual modalities, hypothesizing that first impressions were at least partly formed based on nonverbal behavior. The linear relationships between nonverbal behavior and the organizational constructs of hirability and personality were examined in a correlation analysis. Last, we conducted inference experiments to assess the amount of variance that could be explained by automatically extracted nonverbal cues, and analyzed the predictive validity of feature groups. Results showed that most constructs could be inferred significantly more accurately than the baseline-average model, with up to 27% of the variance explained for extraversion, and over 20% for social and communication skills.

Figure 7.1 displays an overview of this chapter. In Section 7.1, we present the method used for the data collection of video resumes from YouTube. In Section 7.2, we present and analyze the crowdsourcing experiments designed to collect annotations of demographics, skills, and first impressions of personality and hirability. In Section 7.3, we analyze the structure of perceived skills in a clustering experiment. In Section 7.4, we present the nonverbal cues extracted from the audio and visual modalities. In Section 7.5, the linear relationships between nonverbal cues and the organizational constructs of personality and hirability are investigated in a correlation analysis. In Section 7.6, we present the experiments conducted to infer personality and hirability variables using nonverbal cues as predictors. Finally, we conclude in Section 7.7.

The material in this chapter has not been previously published.

7.1 Data collection

In this section, we describe the method used to collect a dataset of conversational video resumes from YouTube. Conversational videos are videos where the person mostly speaks in front of a camera, in a sensor setting similar to video blogs [28]. Video resumes are not necessarily conversational: for instance, *creative* video resumes are often used by graphic designers and artists to display their skills, resulting in stop-motion videos or graphical animations; *Pow-*



(1) (2) (3)

Figure 7.2 – Examples of three types of video resumes: (1) Conversational, (2) Creative, and (3) PowerPoint. We focus on conversational video resumes.

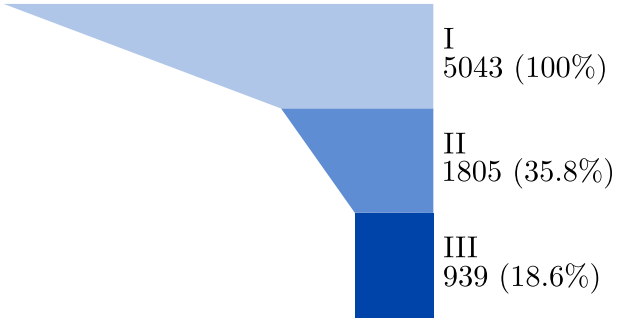


Figure 7.3 – Funnel chart illustrating the collection of English-speaking conversational video resumes. (I) 5043 videos were downloaded from YouTube. (II) 1805 were conversational video resumes upon manual inspection. (III) 939 conversational video resumes were in English; they form the video resume dataset. The face has been blurred due to privacy reasons.

erPoint video resumes constitute another type of video resumes, where slides are displayed and an off-screen voice comments them. Figure 7.2 displays examples of conversational, creative and PowerPoint video resumes. In this study, we solely focused on conversational video resumes because they constitute the setting where behavior can play an important role in the formation of first impressions. Additionally, we concentrated on English-speaking conversational video resumes to ensure that raters understand the verbal content of the videos. Figure 7.3 displays the funnel chart as a summary of the data collection process.

Potential video resumes were queried using the public YouTube Data API [14] to search by keyword and by channels. Keywords were selected by manually browsing through YouTube, and channels found to be specialized in video resumes were also used; Table 7.1 displays the keywords and channel IDs used. In total, 5043 unique video IDs were returned by these queries, and each video was downloaded using youtube-dl [15], a command-line program designed to download videos from online video platforms. Videos were downloaded at a maximum of VGA resolution (640 × 480).

The downloaded videos were manually filtered to only keep conversational video resumes. To this end, we used a custom-built script to view and keep/discard videos based on keyboard shortcuts. The filtering was completed by the PhD candidate, and took in total about 16 hours (approximately 11 seconds per video). From the 5043 videos downloaded, 1805 videos were

Table 7.1 – List of keywords and channels used to query YouTube for video IDs.

Keywords:	video curriculum, video cv, video curriculum vitae, vatel cv, video internship cv, video internship resume, digital cv, video resume
Channels:	vatellosangeles, CURRICULUM VIDEO - votre CV VIDEO en studio, Zookel - video Cvs, impressionsDWREC, Video Resume Now

rated as conversational video resumes.

Crowdourced annotations of language were completed for all 1805 conversational video resumes using Amazon Mechanical Turk [1], among other variables (see Section 7.2). Out of the 1805 videos, 939 were rated as English-speaking conversational video resumes. These 939 video resumes form the video resume dataset used in the next sections of this chapter.

7.2 Crowdsourced annotations of video resumes

One of the objectives of this chapter is to understand the demographics of job-seekers uploading video resumes on YouTube. For instance, what is the age, ethnicity, and gender of the people in the dataset? What type of jobs are they applying to? To the best of our knowledge, such statistics on video resumes have not been reported previously. Another objective of this chapter is to assess the reliability of unacquainted naïve judges in the formation of first impressions on conversational video resumes. For instance, what skills can be consistently annotated by naïve raters? Can high-level organizational social constructs such as personality or hirability be reliably rated? To address these questions, Amazon Mechanical Turk (MTurk) [1] represents an affordable, fast, and fully scalable method to collect annotations of videos. In this section, we present the crowdsourced annotations of video resumes collected using MTurk. Although traditionally used for labeling images or providing product descriptions, related work has shown that agreement among MTurk raters could be found on first impressions: research studies have successfully collected MTurk impressions of urban places using pictures as stimulus [118], as well as personality from conversational video blogs [29].

7.2.1 Method

Four Human Intelligence Tasks (HITs), *i.e.* individual tasks that MTurk workers work on, were designed to annotate (1) basic facts (gender, language, audio/video quality), (2) demographics (age, ethnicity, seniority level, job categories), (3) perceived skills, and (4) first impressions of hirability and personality. In order to ensure that MTurk workers watched a part of the video (*i.e.*, to prevent spammers), the videos had to be watched for a minimal duration (15 seconds for the Basic Facts HIT, 45 seconds for the other HITs) before being able to start answering the questions. For all HITs, all videos were annotated by 5 MTurk raters for a price ranging

between 15¢ and 20¢ per HIT. To be able to work on the HITs, MTurk workers needed to have over 95% positive feedback in their MTurk previous jobs. In order to minimize cultural biases, we controlled for the origin of MTurk raters: they had to be located in the U.S. As a possible future research direction, cultural biases in hirability impressions could be studied by using pools of raters from various origins.

Crowdsourced annotations given by each rater were aggregated to obtain one score per video and per variable. For categorical variables, the aggregation was obtained by majority voting. Likert variables were considered as continuous (even if strictly speaking this was not the case), and the aggregation was obtained by taking the average value over all judges.

To assess the reliability of each annotated variable in the absence of ground truth, we used inter-rater agreement metrics. For categorical variables, (*e.g.* ethnicity, gender), we used Fleiss' Kappa to assess inter-rater reliability. This statistic is similar to Cohen's Kappa, but was designed for any number of raters giving categorical ratings. Fleiss' Kappa assumes a fixed number of raters, but items are not assumed to be rated by the same individuals, which makes this statistic suitable for the task at hand. Fleiss' Kappa can be interpreted as the degree of agreement between the raters above the level of agreement expected by chance [52]. In psychology, $\kappa \in [.40,.60]$ is considered as moderate agreement, while $\kappa \in [.60,.80]$ is substantial and $\kappa > .80$ is almost perfect [80]. However, Fleiss' Kappa is sensitive to unbalanced categories and depends on the number of categories, therefore the interpretation of the Kappa values should be handled with caution. For numerical variables, we used one of the Intraclass Correlation Coefficients (ICC) to assess the level of agreement among judges, as they are commonly used in psychology for this task [125]. Because each target (*i.e.*, video resume) is rated by a different set of judges, only $ICC(1,1)$ and $ICC(1,k)$ can be used [125]. $ICC(1,1)$ measures the extent to which two judges agree with each other, while $ICC(1,k)$ assesses the degree of agreement in rating the targets when the ratings are aggregated across the judges. Because ratings from multiple raters are aggregated by taking the average value, $ICC(1,k)$ is the most suitable metric for our task. Additionally, similarly to Fleiss' Kappa, $ICC(1,k)$ is sensitive to the variance present in the data: for instance, perfect agreement can yield low $ICC(1,k)$ values if the variance is very low. For this reason, we also report the mean and standard deviation in the analyses of inter-rater agreement. Furthermore, $ICC(1,k)$ values can be problematic to interpret as no standard threshold exists to segment between *e.g.* moderate and high agreement. To address this issue, we compared our results with the literature investigating the agreement of judges on related social dimensions. Despite a lack of a clear-cut interpretation for $ICC(1,k)$ values, we used a threshold of $ICC(1,k) = .50$ as a cut-off between low and high inter-rater agreement.

7.2.2 Basic Facts HIT

This HIT was used in Section 7.1 for the annotation of English-speaking video resumes. In addition to gender (male *vs.* female) and language (Engish *vs.* other), we designed the HIT to

verify whether the selected videos were conversational video resumes, using two categorical variables, namely conversational (conversational *vs.* non-conversational *vs.* not sure) and video resume (video resume *vs.* other *vs.* not sure). To this end, we provided MTurk raters with clear guidelines about what was a conversational video and a video resume, with examples and counter examples. Three questions were asked about the perception of the video quality: video quality, audio quality, and overall quality, which were answered on a 5-point Likert scale. In total, all 1805 videos saved after the manual filtering step (Section 7.1) were annotated.

As one would expect, the level of agreement was high for gender (κ = .97) and language (κ = .92). Lower values were found for labelings of conversational videos (κ = .11) and video resumes (κ = .13), but this can be explained by the fact that the aggregated class distribution was strongly unbalanced (94.5% *yes* for conversational and 99.5% *yes* for video resumes). Moreover, 60.2% and 92.1% of the videos had full agreement for conversational and video resume, respectively, which shows that agreement on these variables was high despite the low κ values. Finally, the level of agreement for audio, video, and overall quality was high ($ICC(1,k) \in [.75,.77]$). These results suggest that the HITs were conscientiously completed by MTurk workers, and that the manual selection of conversational video resume was consistent with the MTurk annotations.

7.2.3 Demographics HIT

The demographics HIT was designed to collect additional information beyond gender and language, and included the seniority level, the age, the ethnicity, the language proficiency, and the job category of the job seekers depicted in the vido resumes. the motivation to collect such annotations was to understand the demographics of job-seekers recording video resumes. Nine types of job categories were used based on the American Time Use Survey (ATUS) [2], a tool developed by the U.S. Bureau of Labor Statistics to measure the amount of time people spend doing various activities, such as paid work, childcare, volunteering, and socializing. ATUS reports statistics based on a total number of over 140000 U.S. employees based on occupations, therefore we believe that it constitutes a reliable categorization system for our task. To our knowledge, this use of ATUS is novel. The list and descriptions of the job categories used for this HIT are shown in Table 7.2. The descriptions for each job category were provided in the HIT, and each job category was rated on a 5-point Likert scale. All 939 videos of the video resume dataset were annotated.

As one would expect, the level of agreement among judges was high for age ($ICC(1,k)$ = .87), language proficiency ($ICC(1,k)$ = .89), ethnicity (κ = .66), and seniority ($ICC(1,k)$ = .59). Table 7.3 displays the interrater agreement for job categories. These results show that most job categories were reliably annotated with $ICC(1,k)$ values ranging from .58 to .85, with the exception of production ($ICC(1,k)$ = .31) and office ($ICC(1,k)$ = .44). These results suggest that MTurk raters were reliable in the task of annotating the type of position job-seekers were applying to, and that the annotation of job categories in a crowdsourcing setting is a feasible

Table 7.2 – Descriptions of the job categories annotated in Section 7.2.3.

Job category	Description
Construction	Construction, extraction, installation, maintenance, and repair.
Creative	Design, advertisement, music, and arts.
Healthcare	Nurses, doctors, and personal care.
Hospitality	Accommodation, restaurants and bars, travel and tourism.
Management	Management, business, and financial occupations.
Office	Office and administrative support.
Production	Production, transportation, and material moving.
Professional	Computer, engineering, or scientific occupations.
Sales	Sales-related occupations.

Table 7.3 – Inter-rater agreement for job categories, using the Intraclass Correlation Coefficient ($N_{videos} = 939$, $N_{raters} = 5$).

	ICC(1,k)	Mean	STD		ICC(1,k)	Mean	STD
Construction	0.62	1.40	0.60	Office	0.44	2.44	0.82
Creative	0.74	2.10	1.02	Production	0.31	1.33	0.43
Healthcare	0.70	1.32	0.57	Professional	0.58	3.19	0.97
Hospitality	0.85	2.06	1.18	Sales	0.63	2.01	0.85
Management	0.72	2.79	1.06				

task.

7.2.4 Skills HIT

This HIT was designed to assess the reliability of naïve judge in the task of rating work-related skills, rather than job domains. Note that by skills we denote an umbrella term including actual skills, traits, and states. We used an initial list of 25 skills known to be often assessed in paper resumes and during job interviews [48]. The list of skills and their descriptions are shown in Table 7.4. Skills descriptions were provided in the HIT. MTurk raters were asked to answer the question "I see the person as...", and had to rate the person's skill on a 5-point Likert scale. A subset of 200 randomly sampled videos from the video resume dataset was annotated.

For these perceived skills, intraclass correlation coefficients are shown in Table 7.5. Generally, MTurk workers were reliable in rating perceived skills, with 21 out of 25 skills with $ICC(1,k) > .50$. Some of the skills (persuasive, creative, professional, clear, confident, enthusiastic) had reliabilities greater than .65, which is comparable to the ones obtained for job categories. Honest, open-minded, empathic, and stressed were observed to have low interrater agreement ($ICC(1,k) < .50$), suggesting that these skills were not evident to rate, or that the setting did not elicit clear displays of such skills. The low interrater agreement obtained for stressed can be explained by the fact that a high value indicated a negative perception, which was the opposite for all other skills; this could have confused the raters. Additionally, stress has been previously reported as a variable with relatively low interrater agreement [120].

Table 7.4 – Descriptions of the skills used annotated in Section 7.2.4.

Skill	Description
Clear	Easy to perceive, understand, or interpret.
Communicative	Willing, eager, or able to talk or impart information.
Competent	Having the necessary ability, knowledge, or skills to do something successfully.
Concise	Giving a lot of information clearly in a few words; brief but comprehensive.
Confident	Feeling or showing confidence in oneself or one's abilities or qualities.
Creative	Involving the use of the imagination or original ideas to create something.
Dedicated	Devoted to a task or purpose.
Empathic	Able to understand and share the feelings of others.
Enthusiastic	Having or showing intense and eager enjoyment, interest, or approval.
Friendly	kind and pleasant.
Funny	Causing laughter or amusement; humorous.
Hard-Working	Tending to work with energy and commitment; diligent.
Honest	Free of deceit; truthful and sincere.
Independent	Not relying on others for aid and support.
Intelligent	Having or showing intelligence, especially of a high level.
Leader	A person who has the ability to lead or command a group or an organization.
Motivated	Enthusiastic and determined to achieve success.
Open-Minded	Willing to consider new ideas; unprejudiced.
Organized	Able to plan one's activities efficiently.
Persuasive	Good at persuading someone to do or believe something.
Positive	Constructive, optimistic, or confident.
Professional	Worthy of or appropriate to a professional person; competent, skillful, or assured.
Reliable	Consistently good in quality or performance; able to be trusted.
Sociable	Willing to talk and engage in activities with other people; friendly.
Stressed	Mentally tensed and worried.

Table 7.5 – Inter-rater agreement for perceived skills, using the Intraclass Correlation Coefficient ($N_{videos} = 192$, $N_{raters} = 5$).

	ICC(1,k)	Mean	STD		ICC(1,k)	Mean	STD
Clear	0.74	3.50	0.82	Independent	0.55	4.00	0.57
Communicative	0.68	3.84	0.67	Intelligent	0.52	4.02	0.55
Competent	0.62	3.99	0.63	Leader	0.62	3.32	0.72
Concise	0.59	3.56	0.71	Motivated	0.57	4.06	0.57
Confident	0.67	3.88	0.69	Open-Minded	0.38	3.73	0.48
Creative	0.65	3.35	0.69	Organized	0.63	3.81	0.62
Dedicated	0.51	4.00	0.55	Persuasive	0.67	3.33	0.71
Empathic	0.49	3.33	0.52	Positive	0.61	3.99	0.56
Enthusiastic	0.68	3.64	0.72	Professional	0.70	3.86	0.75
Friendly	0.63	3.91	0.59	Reliable	0.60	4.00	0.56
Funny	0.55	2.60	0.68	Sociable	0.65	3.70	0.65
Hard-Working	0.55	4.04	0.56	Stressed	0.25	2.05	0.54
Honest	0.36	4.02	0.44				

7.2.5 First impressions HIT

One of the main objectives of this chapter is to investigate whether reliable first impressions of hirability can be made on video resumes, and to understand the basis on which these impressions are made. To this end, we designed a HIT aiming at collecting hirability, high-level skills, and personality impressions. In this HIT, we asked MTurk raters to give their first impressions on a video resume. Specifically, MTurkers were asked to rate two variables for the general first impression (overall hirability and overall first impression), three high-level skills derived from the clustering of skills (professional, social, and communication skills, see Section 7.3 for details), and ten items for perceived personality. For personality, we used the Big-Five model which represents personality at its highest level of abstraction, and which consists of five factors, namely extraversion, neuroticism, openness to experience, agreeableness, and conscientiousness [59]. To assess personality, we used the standard Ten Item Personality Inventory (TIPI) questionnaire consisting of ten items, two per dimension [59]. Each question was answered on a five-point Likert scale. All 939 videos of the video resume dataset were annotated.

Figure 7.6 displays the inter-rater agreement for the first impressions HIT. Hirability first impressions (overall hirability, overall impression, and professional, social, and communication skills) were observed to have moderate to high inter-rater agreement with $ICC(1,k) \in [.59,.64]$. These values are comparable to the ones obtained for low-level skills, and somewhat lower than the job categories, but still demonstrate that reliable first impressions of hirability from naïve raters can be obtained from video resumes. For personality, only extraversion was observed to be consistently rated ($ICC(1,k) = .64$). Openness to experience showed moderate to low agreement ($ICC(1,k) = .46$), while agreeableness, conscientiousness, and neuroticism had low inter-rater agreement ($ICC(1,k) \in [.18,.38]$). In comparison with other works investigating the reliability of personality impressions, several observations can be noted. First, extraversion was the trait that achieved the highest level of agreement, which has been repeatedly observed in related works [29, 60]. Second, overall agreement on personality impressions were observed to be lower than the ones obtained from video blogs [29]. In order to understand whether this relatively low inter-rater agreement was due to how the first impression HIT was designed or from the setting itself, we collected a second round of MTurk annotations on a subset of the data ($N = 200$), where only personality traits were rated using the TIPI questionnaire [59]. Inter-rater agreement results resulting from this experiment were observed to be very close to the ones reported in Table 7.6 and are not shown here in detail. This finding suggests that with the exception of extraversion (and openness to experience, to a lesser extent), personality impressions were difficult to consistently rate from video resumes. This could be explained by the fact that personality traits other than extraversion and openness to experience might not be expressed in this setting. This hypothesis would have to be validated as part of future work.

Table 7.6 – Inter-rater agreement for annotations from the first impression HIT, using the Intraclass Correlation Coefficient ($N_{videos} = 939$, $N_{raters} = 5$).

	ICC(1,k)	Mean	STD		ICC(1,k)	Mean	STD
Overall Impression	0.59	3.70	0.62	Extraversion	0.64	3.46	0.61
Overall Hirability	0.61	3.72	0.62	Agreeableness	0.27	3.71	0.39
Professional Skills	0.59	3.76	0.60	Conscientiousness	0.38	3.91	0.46
Social Skills	0.57	3.67	0.63	Neuroticism	0.18	2.21	0.40
Communication Skills	0.64	3.71	0.69	Openness	0.46	3.51	0.51

7.2.6 Demographics of video resumes

Figure 7.4 displays the demographics (gender, age, ethnicity, job type, seniority level, and duration) of the video resume dataset. We first observe that the great majority of the population was young (59% were under 25 and 93% were under 35) and either looking for an internship position (37%) or a junior position (41%). Gender was unbalanced as there were approximately twice as many males (66%) as females (34%). In terms of ethnicity, the video resume dataset was mainly populated by Caucasian (36%) and Indian (32%) job-seekers. One possible hypothesis to explain this finding stems from the bias generated by the decision of keeping only English-speaking video resumes. For instance, many Spanish-speaking video resumes were present in the set prior to filtering out non-English video resumes. For job categories, the most represented job category was professional (computer, engineering, or science occupations) accounting for 39% of the dataset, followed by management (21%), hospitality (16%), and creative (12%). We hypothesize the creative job category to be under-represented in this dataset because only conversational video resumes were considered; if the creative video resumes (see Figure 7.2) were also included, this job category would likely be more represented. In terms of duration, the majority of the video resumes are relatively short, with 62% lasting less than 150 seconds with the main mode situated between 50 and 100 seconds. The distribution was long-tail like, and the maximum video resume duration in the dataset is 818 seconds.

Demographics varied across job categories. The professional category (engineers, scientists, and computer scientists) was populated by 77% males, 50% Indians, and 26% Caucasians, suggesting that the cliché of male Indian engineers and computer scientists was observed in the video resume dataset. One missing part in our dataset is the geographic distribution of uploads. this could explain possible biases related to the home location of the applicants. For hospitality and creative applicants, the gender was more balanced (48% and 57% males, respectively) and these categories were dominated by Caucasian applicants (28% and 64%, respectively). In terms of age and seniority distributions, applicants in the management category were older and were applying for more senior positions, and no major difference between the professional, hospitality, and creative job categories was observed.

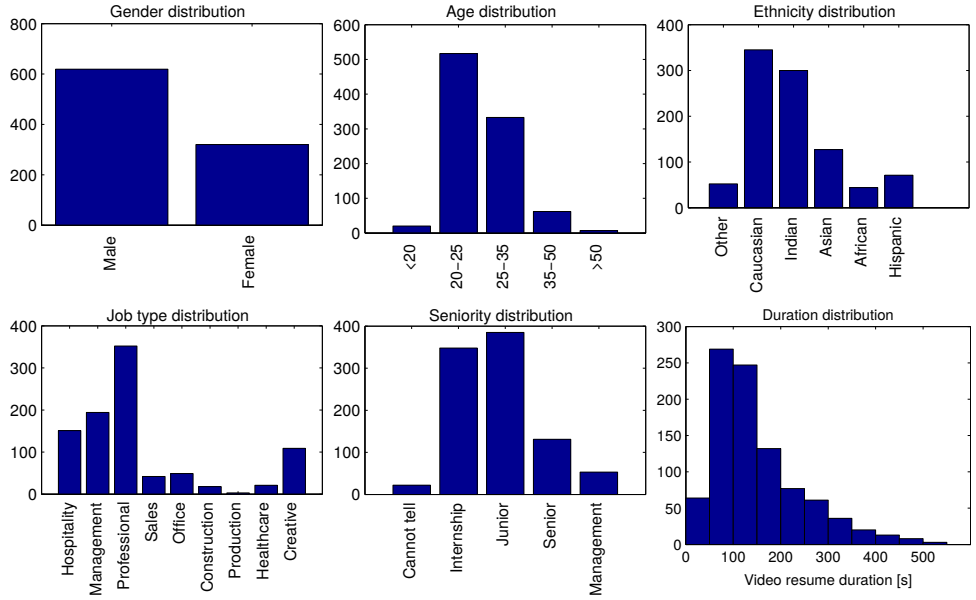


Figure 7.4 – Demographics (gender, age, ethnicity, job type, seniority, and video resume duration) of the video resume dataset ($N = 939$).

7.3 Clustering of skills

To understand the structure of perceived job-related skills, we conducted a clustering analysis of the skills annotated in the Skills HIT on 200 video resumes (Section 7.2.4). The values were first pre-processed: skills with low inter-rater agreement (($ICC(1,k) < 0.5$) were removed, and each variable was standardized such that it had zero mean and unity variance. This left a set of 21 skills.

We then conducted principal component analysis (PCA) on the skills variables, which constitutes a projection onto an orthogonal space of lower dimension. It learns the linear transfor-

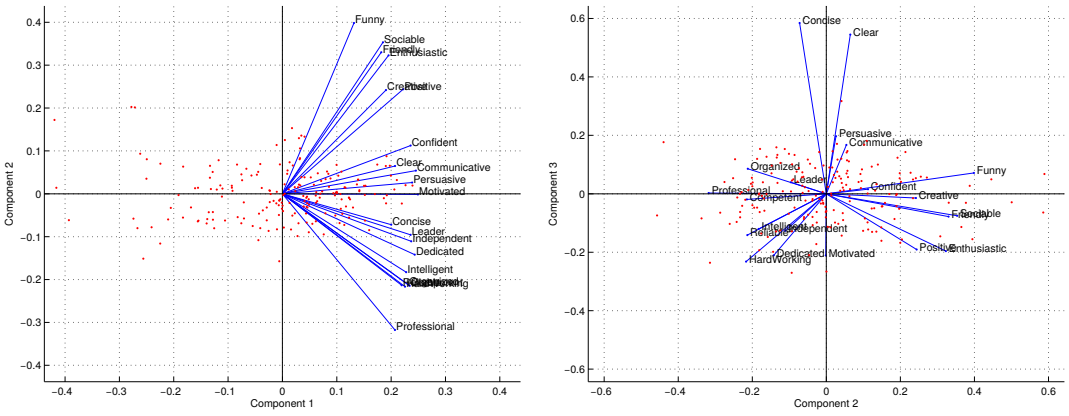


Figure 7.5 – First three principal components of the principal component analysis on perceived skills ($N = 200$), accounting for 82% of the variance.

mation such that the variance of the projected points is maximized [74]. Figure 7.5 displays the video resume data points as well as the original variables projected onto the coordinate system of the three first principal components, accounting for 82% of the variance in the data. One can observe that the first component, accounting for 64% of the variance, was not very distinctive of skills as they were all positive, but was evocative of the fact that video resumes were rated mainly on a good/bad basis. In psychology, this effect is known as the halo effect, where global evaluations tend to induce altered evaluations of other attributes [100]. The second principal component, accounting for 14% of the variance, seemed to encompass two different social concepts. On one hand, the variables of funny, sociable, friendly, enthusiastic, creative, positive seemed to be part of a *social* high-level label; on the other hand, *professional* variables seemed to form a cluster including competent, organized, professional, reliable, hard-working, etc. The third component accounted for 4% of the variance, and seemed to enclose *communication* skills: concise, clear, persuasive, and communicative.

K-means clustering of the original variables was performed in the principal component space. K-means is a simple iterative clustering algorithm consisting of two steps: first, data points are assigned to clusters based on the distance to cluster centers (defined by a distance metric); second, cluster centers are updated by taking the average of all points assigned to the cluster; these two steps are repeated until convergence is achieved [32]. The number of clusters K is the only parameter to be specified. We used the Euclidean distance as distance measure, but PCA coordinates were multiplied by the square-root of the eigenvalues corresponding to each component (thus accounting for the standard deviation explained by each principal component) prior to completing K-means clustering. Experiments with $K \in [2,7]$ were completed, and $K = 3$ appeared to be a subjectively optimal choice.

Figure 7.6 displays the three obtained clusters of perceived skills, where colors denote correlation coefficients (warm is positive, cold is negative). A first dense cluster including organized, motivated, independent, dedicated, intelligent, competent, professional, hard-working, leader, and reliable was observed, and was labeled as *professional* skills. The second cluster encompassed creative, friendly, enthusiastic, positive, funny, and sociable, and was labeled as *social* skills. The last cluster included persuasive, clear, concise, communicative, and confident, and was labeled as *communication* skills. As a verification step, factor analysis was also completed on the skills, and similar clusters were found.

The three obtained high-level skills of this Section were then annotated (s already discussed in Section 7.2.5) for all video resumes and used as hirability variable for the correlation and inference analyses (Sections 7.5 and 7.6).

7.4 Extraction of behavioral cues

One of the objectives of this chapter is to investigate the use of automatically extracted behavioral cues for the inference of first impressions in conversational video resumes. To this end, we used existing methods to extract simple behavioral cues from both the audio and

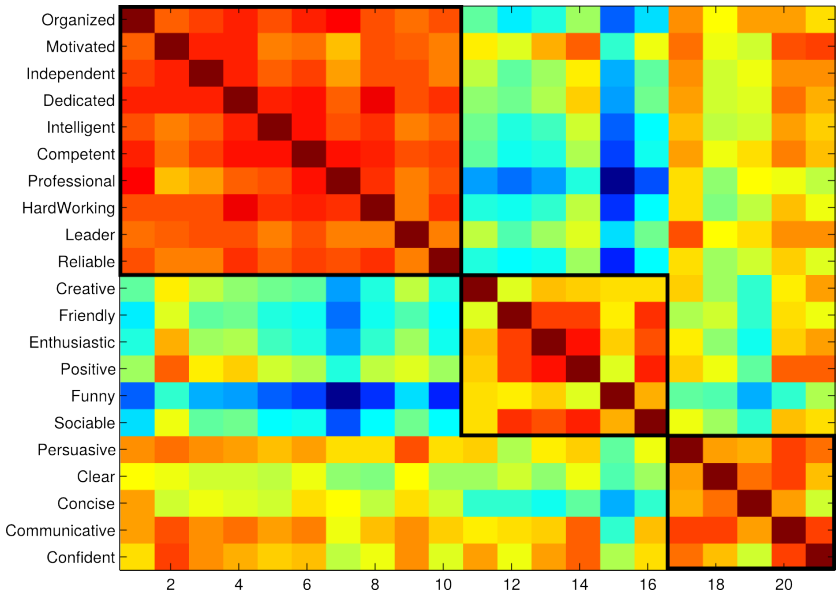


Figure 7.6 – K-means clustering on skills ($N = 200$), with $K = 3$. Warm and cold colors denote positive and negative correlation coefficients, respectively.

visual modalities. We address the problem from a nonverbal perspective where the feature extraction is fully automated. The choice of nonverbal cues to be extracted was based on the nonverbal communication literature [78] and the existence of available processing methods from our previous work described in earlier chapters. Because job-seekers do not present themselves in front of another person, conversational video resumes cannot be considered as face-to-face social interaction as such. Nevertheless, they consist of a person delivering a message in a natural manner, and we hypothesize that the findings in social psychology stating that nonverbal behavior influences the way people are perceived still apply [78]. As discussed in Chapter 2, recent related work in social computing have observed that nonverbal cues could be predictive of social constructs in non-face-to-face interactions such as video blogs (*e.g.*, [29]) or human-computer interfaces (*e.g.*, [25]).

In comparison with datasets recorded in laboratory settings using high-end audio and video sensors and optimal illumination like the ones used earlier in this dissertation, user-generated online videos are challenging to process. The presence of music, improper illumination, low-quality audio recording, high compression rates, low video resolution, non-fixed cameras, text displays and overlays, multiple people on the video, and unexpected user behavior count among the many challenges for automatically processing online video resumes. As a result, simple and robust extraction methods were preferred over fine-grain but more error-prone algorithms. Furthermore, the set of extracted behavioral features was assumed to be inherently noisy to the challenging nature of online video resumes.

In this section, we present the behavioral cues extracted from the audio and visual modalities, which will then be used as a feature representation of the video resumes for the inference of

social variables.

7.4.1 Audio cues

Turn-based and prosodic features were extracted from the video resumes, based on the MIT Media Lab Speech Extraction Code [7].

Speaking turns

We defined speaking turn events as a sequence of audio frames where speech was detected. Speaking turn events were characterized by their starting and ending times. Statistics on the speaking turn durations were computed for each video resume: mean, median, standard deviation, minimum, maximum, and quartiles, as well as the number of turns and the ratio of speaking time.

Prosody

- **Voiced rate**. Cues related to speech rate were derived from the voiced/non-voiced segmentations obtained from the Speech Feature Extraction Code [7]. Voiced events were defined as a sequence of audio frames classified as *voiced*, and were characterized by their starting and ending times. Statistics on the durations of voiced segments were computed: mean, median, standard deviation, minimum, maximum, quartiles, and entropy, as well the ratio of voiced time over speaking time.
- **Speech energy**. Cues related to speech energy were obtained by squaring the value of each audio sample classified as *speaking*. Statistics over the time-series were computed: mean, median, standard deviation, minimum, maximum, quartiles, and entropy.
- **Pitch**. Cues related to pitch were derived from the pitch values obtained from the Speech Feature Extraction Code [7]. Statistics of the pitch values were computed: mean, median, standard deviation, minimum, maximum, quartiles, and entropy.

7.4.2 Visual cues

The visual modality of the video resumes was characterized by nonverbal cues capturing proximity, looking activity, head motion, and body activity. These visual cues were successfully used to characterize people in a setting similar to video resumes, namely video blogs [29], and have the main advantage to be easy to extract.

Proximity, looking activity, and head motion

Proximity, looking activity, and head motion cues were derived from the output of a standard frontal face detection system based on the object detection method proposed by Viola and

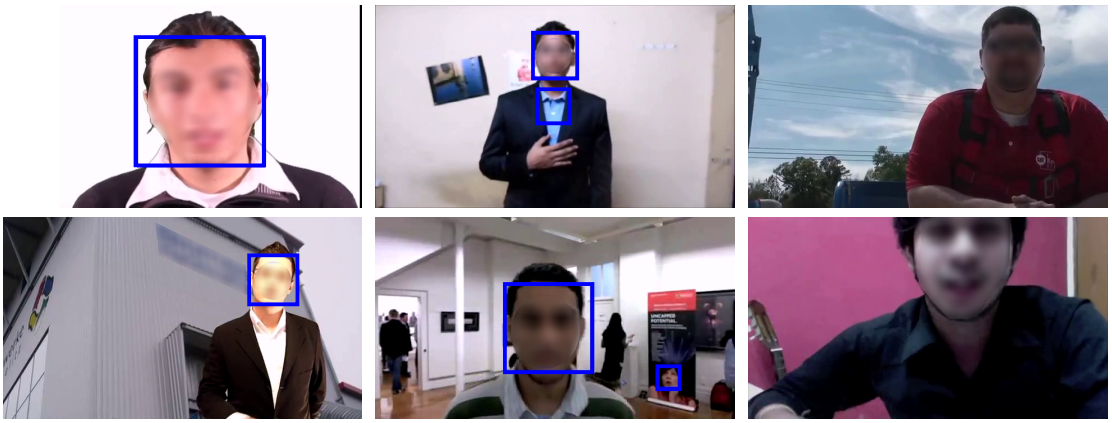
Jones [135]. Over more elaborate tracking algorithms, this method has the advantages of being robust, simple to implement, and not requiring any initialization or parameter tuning. For each frame, the algorithm extracted the bounding box of each detected frontal face, characterized by its position and size. The frontal face detection method was not systematically evaluated, but the output of the algorithm was manually inspected for a subset of 100 randomly sampled video resumes, and in most cases the faces were correctly detected. Among the observed errors, intermittent detection within a video, or false positives detected for a short duration were the most frequent. In some rare cases, no face could be detected for the whole duration of the video, which was generally due to improper framing and excessively bad illumination and/or overall video quality; Figure 7.7 (3) displays two examples of non-detections.

To account for multiple face detections due *e.g.* to the presence of pictures including faces or other objects with face-like appearance (see Figure 7.7 (2)), we hypothesized that the job seeker's face corresponded to the largest and highest bounding box, and discarded all other detected faces based on a simple cost function. Intermittent detections were also handled by removing each sequence of detected faces that lasted less than half a second, and by assigning sequences of non-detections shorter than half a second to the closest detected bounding box.

- **Proximity**. We used the size of the detected face bounding boxes as a proxy for the distance between the person's face and the camera. To account for the variability of video resolution across video resumes, the size of the bounding boxes was divided by the overall resolution. Statistics over the size of the bounding boxes were computed: mean, median, standard deviation, minimum, maximum, quartiles, and entropy.
- **Head motion**. As a proxy for head motion, we used the displacement of the detected bounding box centers of two consecutive frames, and computed statistics for the horizontal and vertical displacements, as well as for the magnitude: mean, median, standard deviation, minimum, maximum, quartiles, and entropy.
- **Looking activity**. We used detected frontal faces as a proxy for looking at the camera. We defined looking events as a sequence of frames where a frontal face was detected, and each event was characterized by its starting and ending time. To encode looking activity patterns, we computed statistics on the duration of looking events: mean, median, standard deviation, minimum, maximum, and quartiles, as well as the ratio between the number of looking frames and the total number of frames.

Overall visual motion

The overall visual motion of a job-seeker in a video resume can indicate his level of kinetic expressiveness. To measure the overall movement, we used the Weighted Motion Energy Image (WMEI) descriptor [27] used in earlier chapters of the thesis. As descriptors for the overall visual motion, we computed statistics on the WMEIs: mean, median, standard deviation, minimum, maximum, entropy, quartiles, and center of gravity.



(1) (2) (3)

Figure 7.7 – Examples of frontal face detections: (1) correct detections, (2) multiple detections, (3) missed detections. Faces have been blurred due to privacy reasons.

7.4.3 Video statistics

Basic video statistics were extracted using ffmpeg [3]. These statistics include the size, duration, bitrate, audio sampling rate, horizontal, vertical, and overall resolution, and the framerate.

7.4.4 Feature pre-processing

Prior to carrying out further analyses, nonverbal cues were pre-processed. Data for which no face or no speech was detected were discarded. In total, 882 videos were kept for analysis (57 removed). Highly skewed features ($skew > 1$) were log-transformed ($x' = log(1+x)$ where x' and x denote the transformed and original features, respectively). Finally, all nonverbal cues were standardized, such that each cue had zero mean and unity variance.

7.5 Correlation analysis

To understand the existing linear relationships between nonverbal behavior, personality, and hirability impressions, we performed a correlation analysis. To this end, we use the Big-Five personality variables (extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience) and the five hirability variables (overall impression, overall hirability, professional skills, social skills, and communication skills), obtained from the first impressions HIT described in Section 7.2.5.

7.5.1 Personality and hirability

Table 7.7 displays the pairwise correlations between the Big-Five personality variables and the hirability variables. High correlation ($r = .69$) was observed among the personality traits

Table 7.7 – Pairwise correlations between the personality and hirability variables, using Pearson's correlation coefficient. All values are statistically significant ($p < 10^{-12}$), $N = 939$.

	1	2	3	4	5	6	7	8	9	10
1. Extraversion		0.27	0.25	-0.33	0.69	0.55	0.51	0.31	0.66	0.62
2. Agreeableness			0.39	-0.43	0.43	0.35	0.35	0.24	0.40	0.32
3. Conscientiousness				-0.65	0.40	0.58	0.63	0.62	0.46	0.45
4. Neuroticism					-0.45	-0.49	-0.51	-0.45	-0.47	-0.46
5. Openness						0.58	0.54	0.37	0.66	0.63
6. Ov. Impression							0.87	0.75	0.79	0.81
7. Ov. Hirability								0.80	0.77	0.78
8. Professional Sk.									0.59	0.63
9. Social Sk.										0.84
10. Communication Sk.										

of extraversion and openness to experience, suggesting that job seekers who were rated high on the trait of extraversion were likely to get high scores for openness to experience. The neuroticism trait was observed to be negatively correlated with all other variables, particularly with conscientiousness ($r = -.65$). Other correlations among personality impressions were lower, with $|r| \in [.25, .45]$.

Hirability variables were found to be strongly inter-correlated ($r \in [.59, .81]$), and the strongest relationship was observed between overall impression and overall hirability ($r = .87$). Social and communicative skills were also strongly correlated ($r = .84$), and interestingly, were also highly correlated with the personality traits of extraversion and openness to experience ($r \in [.62, .66]$). As for the variable of professional skills, it was associated to the personality variable of conscientiousness ($r = .62$).

7.5.2 Nonverbal behavior and personality

Table 7.8 displays Pearson's pairwise correlation coefficients between the nonverbal cues extracted in Section 7.4 and crowdsourced personality impressions, where only cues with $p < 10^{-3}$ are shown. We first observe that the traits of extraversion and openness to experience were the only personality variables with a large number of low to moderate correlated cues. Interestingly, extraversion and openness were also the two traits with acceptable inter-rater agreement ($ICC_{extra}(1,k) = .64$ and $ICC_{open}(1,k) = .46$, see Table 7.6). Given the low inter-rater agreement of agreeableness, conscientiousness, and neuroticism, it is not surprising to observe low correlation values. The sets of nonverbal cues correlated with extraversion and openness to experience were observed to be very similar and have comparable correlation coefficients. This finding can be explained by the fact that these two traits were highly correlated ($r = .69$, see Table 7.7).

Surprisingly, only a few vocal cues were correlated with extraversion and openness to experience. Apparently, this finding contradicts previous studies where vocal cues were found to be predictive of personality traits in general, and extraversion in particular, independently

of the type of setting [29, 82]. Furthermore, the literature on nonverbal communication has established that extraverted individuals are mainly characterized by a large amount of speech, voice modulation, and energy [78]. One possible hypothesis to explain this finding could be that the errors in the vocal cue extraction process generated noisy features; this however needs to be investigated in future work.

Looking turns were observed to be correlated with extraversion and openness to experience. In particular, the correlation coefficients for the number of looking turns was positive, and they were negative for looking turn durations, which was also observed by Biel *et al.* in a study on video blogs [29]. This could suggest that job-seekers who did not consistently look at the camera were perceived as more extraverted and more open, but contradicts the nonverbal behavior literature that established that gaze was associated with extraversion [78]. Looking events were defined as segments where a frontal face was detected, and looking turn breaks could also result from head movements, inserts of short non-conversational video snippets (*e.g.*, the job-seeker showing himself doing something), or detection failures, instead of gaze avoidance. In any case, these results need to be investigated in greater details.

To a lesser degree, proximity cues were also correlated with extraversion and openness to experience. Large variations of face proximity and smaller face bounding boxes were associated with higher ratings on extraversion and openness to experience. These observations differ but do not contradict previous works, where proximity features were not significantly correlated with these traits in video blogs [29]. This suggests that relative positioning of the candidate has a small yet significant effect. For head motion, a large number of nonverbal cues were significantly correlated with extraversion and openness, with correlation coefficients up to $r = .30$. The linear relationship between head motion cues and the personality dimensions was mainly positive, meaning that job-seekers who displayed larger head motions were perceived as more extraverted and more open. Although head motion cues were not investigated in previous studies focusing on online videos, this observation was confirmed by the nonverbal communication literature, which has shown that extraversion was associated with kinetic expressiveness, including head movements [78]; the relationship between head motion and openness to experience was however not clearly established in the literature [78] but can be explained by the strong correlation observed between extraversion and openness. For overall body motion measured by statistics on weighted motion energy images (WMEI), cues were also found to be correlated with extraversion and openness. Specifically, job-seekers displaying a larger and more diverse visual activity were perceived as more extraverted and open. Again, the nonverbal nonverbal communication literature confirms this observation, as extraverted people are usually kinetically expressive [78]; furthermore, similar results were reported on video blogs [29]. Last, some basic video statistics (bitrate and video size) were positively correlated with extraversion and openness to experience, suggesting that high definition video resumes earned higher ratings for these two traits.

Table 7.8 – Nonverbal cues significantly correlated with at least one personality variable, using Pearson's correlation coefficient ($p < 10^{-3}$, $^{*}p < 10^{-4}$, $^{\dagger}p < 10^{-5}$).

NVB cue	Extra.	Agree.	Consc.	Neuro.	Open.
Vocal cues:					
Speaking ratio		0.12	0.13^{*}		
STD of speaking turn duration					0.11
Median speaking turn duration			0.11	−0.11	
Maximum speaking turn duration	0.13^{*}				$0.16^{\dagger}$
STD of speech pitch		−0.12			
Entropy of speech pitch	$0.15^{\dagger}$				0.14^{*}
Maximum voiced segment duration	0.12				0.11
Looking turns:					
Number of looking turns	$0.16^{\dagger}$				$0.18^{\dagger}$
Average looking turn duration	$-0.20^{\dagger}$				$-0.19^{\dagger}$
Median looking turn duration	$-0.20^{\dagger}$				$-0.19^{\dagger}$
STD of looking turn duration	0.13				
Minimum looking turn duration	$-0.21^{\dagger}$				$-0.20^{\dagger}$
Maximum looking turn duration	$-0.16^{\dagger}$				$-0.15^{\dagger}$
Looking turn duration lower quartile	$-0.21^{\dagger}$				$-0.20^{\dagger}$
Looking turn duration upper quartile	$-0.18^{\dagger}$				$-0.17^{\dagger}$
Looking ratio	-0.14^{*}				$-0.17^{\dagger}$
Proximity cues:					
Average bbox size	−0.12				
STD of bbox size	$0.16^{\dagger}$				$0.20^{\dagger}$
Median bbox size	−0.13				
Minimum bbox size	$-0.23^{\dagger}$				$-0.18^{\dagger}$
Bbox size lower quartile	$-0.15^{\dagger}$				
Head motion:					
Average h. head motion	$0.25^{\dagger}$				$0.23^{\dagger}$
STD of h. head motion	$0.24^{\dagger}$				$0.24^{\dagger}$
Median h. head motion	$0.18^{\dagger}$				$0.15^{\dagger}$
Maximum h. head motion	$0.29^{\dagger}$				$0.30^{\dagger}$
H. head motion lower quartile	$0.21^{\dagger}$				$0.16^{\dagger}$
H. head motion upper quartile	$0.19^{\dagger}$				$0.19^{\dagger}$
Entopy of h. head motion	−0.13				-0.13^{*}
Average v. head motion	$0.20^{\dagger}$				$0.22^{\dagger}$
STD of v. head motion	$0.18^{\dagger}$				$0.20^{\dagger}$
Median v. head motion	$0.16^{\dagger}$				$0.16^{\dagger}$
Maximum v. head motion	$0.25^{\dagger}$				$0.26^{\dagger}$
V. head motion lower quartile	$0.18^{\dagger}$				$0.15^{\dagger}$
V. head motion upper quartile	$0.18^{\dagger}$				$0.20^{\dagger}$
Average head motion	$0.25^{\dagger}$				$0.25^{\dagger}$
STD of head motion	$0.24^{\dagger}$				$0.25^{\dagger}$
Median head motion	$0.20^{\dagger}$				$0.19^{\dagger}$
Maximum head motion	$0.30^{\dagger}$				$0.31^{\dagger}$
Head motion upper quartile	$0.21^{\dagger}$				$0.21^{\dagger}$
Entopy of head motion	-0.13^{*}				-0.14^{*}
Overall motion:					
Average WMEI value	$0.20^{\dagger}$				$0.19^{\dagger}$
Median WMEI value	$0.20^{\dagger}$				$0.18^{\dagger}$
Minimum WMEI value	$0.15^{\dagger}$				$0.17^{\dagger}$
WMEI lower quartile	$0.22^{\dagger}$				$0.23^{\dagger}$
WMEI upper quartile	$0.19^{\dagger}$				$0.17^{\dagger}$
WMEI entropy	$0.25^{\dagger}$				$0.20^{\dagger}$
Video stats:					
Video size					0.13^{*}
Bitrate	$0.18^{\dagger}$				$0.17^{\dagger}$

7.5.3 Nonverbal behavior and hirability

Table 7.9 displays Pearson's pairwise correlation coefficients between the nonverbal cues extracted in Section 7.4 and the hirability variables, where only cues with $p < 10^{-3}$ are displayed. Overall, low to moderate yet statistically significant effects can be observed. First, no vocal cue was correlated with overall hirability, and only three vocal cues were correlated with the overall impression. This is surprising, considering the results obtained on employment interviews in previous chapters, where hirability was observed to be strongly correlated to applicant speaking turn-based and prosodic cues (see Section 5.1). To understand these differences, one should consider the differences of settings between employment interviews and video resumes. First, the job interview dataset was recorded in laboratory settings where the environment was controlled, which positively affected the quality of the recordings in terms of background noise; the conditions were therefore ideal for a clean extraction of vocal cues. In video resumes, the audio quality was not guaranteed (background noise, presence of music, low-quality microphones) and the extraction process might have suffered from it. Second, the structured nature of job interviews allowed to objectively compare the behavior of job applicants, which was particularly important for speaking turn-based cues, whereas in video resumes no such structure existed.

Cues related to head motion were found to be positively correlated with the overall impression and, to a lesser degree, with overall hirability. This suggests that job-seekers who displayed kinetic expression from the head were positively rated. For overall motion, no WMEI cue was correlated with either impression or hirability. Similar results were observed in the job interview corpus (see Section 5.1). For basic video statistics, bitrate was positively correlated with all hirability variables except for professional skills, which indicates that videos of high definition made a better impression on raters.

For social and communication skills, the sets of correlated nonverbal cues were very similar and the correlation values very close. The high correlation found between the two variables ($r = .84$) explains these similarities. Additionally, social and communication skills shared a similar set of correlated cues with the personality variables of extraversion and openness to experience, which can also be explained by the high correlations among these variables ($r \in [.62, .66]$). Compared to other variables, communication skills was correlated to a relatively large number of vocal cues. Specifically, job-seekers who spoke much, and with low energy were perceived as more communicative, and similar observations can be made for social skills to a lesser degree. These observations can be justified by the fact that these skills inherently require speech abilities. The variable of professional skills had a low number of correlated cues, which can either be explained by the hypothesis that the extracted nonverbal cues were too noisy to hold any relevant information, or that professional skills were rated on the basis of other behavioral cues, *e.g.* verbal cues; this result needs to be investigated in greater detail in the future.

Table 7.9 – Nonverbal cues significantly correlated with at least one hirability variable, using Pearson's correlation coefficient ($p < 10^{-3}$, $^{*}p < 10^{-4}$, $^{\dagger}p < 10^{-5}$).

NVB cue	Ov. Impression	Ov. Hirability	Pro. Sk.	Social Sk.	Comm. Sk.
Vocal cues:					
Speaking ratio					0.12
Average speech energy	−0.11			−0.13	$-0.17^{\dagger}$
STD of speech energy	−0.11			−0.13	$-0.17^{\dagger}$
Median speech energy				−0.11	-0.14^{*}
Maximum speech energy					−0.12
Lower quartile of speech energy					-0.13^{*}
Upper quartile of speech energy	−0.11			−0.13	$-0.16^{\dagger}$
Looking turns:					
Average looking turn duration				-0.13^{*}	−0.13
Median looking turn duration				-0.14^{*}	-0.13^{*}
Minimum looking turn duration				-0.14^{*}	-0.14^{*}
Maximum looking turn duration				−0.11	
Looking turn duration lower quartile				-0.14^{*}	-0.14^{*}
Looking turn duration upper quartile				−0.12	−0.11
Proximity cues:					
Average bbox size	−0.11				
Median bbox size	−0.11				
Minimum bbox size	$-0.16^{\dagger}$	−0.12	−0.11	$-0.15^{\dagger}$	−0.11
Bbox size lower quartile	−0.12				
Head motion:					
Average h. head motion	0.11			$0.16^{\dagger}$	$0.16^{\dagger}$
STD of h. head motion	0.13^{*}			$0.16^{\dagger}$	$0.16^{\dagger}$
Median h. head motion	0.12			0.14^{*}	0.13^{*}
Maximum h. head motion	$0.19^{\dagger}$	$0.15^{\dagger}$	0.12	$0.21^{\dagger}$	$0.20^{\dagger}$
H. head motion lower quartile	$0.18^{\dagger}$	$0.16^{\dagger}$		$0.19^{\dagger}$	$0.20^{\dagger}$
H. head motion upper quartile				0.13	0.12
Average v. head motion	0.11			0.14^{*}	0.14^{*}
STD of v. head motion				0.12	0.11
Median v. head motion	$0.15^{\dagger}$	0.12		0.14^{*}	0.14^{*}
Maximum v. head motion	0.14^{*}			$0.17^{\dagger}$	$0.16^{\dagger}$
V. head motion lower quartile	$0.16^{\dagger}$	0.14^{*}		$0.17^{\dagger}$	$0.16^{\dagger}$
V. head motion upper quartile	0.14^{*}	0.12		$0.16^{\dagger}$	$0.15^{\dagger}$
Average head motion	0.12			$0.17^{\dagger}$	$0.16^{\dagger}$
STD of head motion	0.13^{*}			$0.16^{\dagger}$	$0.16^{\dagger}$
Median head motion	0.14^{*}			$0.15^{\dagger}$	$0.15^{\dagger}$
Maximum head motion	$0.19^{\dagger}$	$0.15^{\dagger}$	0.12	$0.21^{\dagger}$	$0.20^{\dagger}$
Head motion lower quartile	$0.15^{\dagger}$	0.12	0.11		
Head motion upper quartile				0.14^{*}	0.13^{*}
Overall motion:					
Average WMEI value				0.13^{*}	
Median WMEI value				0.12	
Minimum WMEI value				0.11	
Maximum WMEI value			0.11		
WMEI lower quartile				$0.15^{\dagger}$	0.11
WMEI entropy				0.11	
WMEI horizontal center of mass			−0.12		
Video stats:					
Bitrate	0.12	0.12		$0.17^{\dagger}$	$0.17^{\dagger}$

7.6 Inference

One of the objectives of this chapter is to investigate the use of automatically extracted nonverbal cues for the prediction of personality and hirability impressions. We defined the inference task as a regression task, where the goal was to infer the exact scores of the variables collected in the first impression HIT (see Section 7.2.5), namely extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience for personality impressions, and overall impression, overall hirability, professional skills, social skills, and communication skills for hirability impressions. We used the same computational framework as the one presented in Chapter 5. Two experiments were completed: first, we assessed the performance of several regression and dimensionality reduction methods (Section 7.6.1); second, we analyzed the predictive validity of feature groups (Section 7.6.2).

We used a 10-fold cross-validation approach for training and testing the regression models. This framework used 90% of the video resumes for training, and kept the 10% remaining for testing. Model parameters were estimated using a 10-fold inner cross-validation approach.

To quantify the performance of the automatic inference models, we used the same performance measures than throughout the thesis, namely the root-mean-square error ($RMSE$) and the coefficient of determination (R^2), as these are two widely used measures in the psychology and social computing (see Section 5.2.2 for details). As the baseline regression model, we took the average score as the predicted value.

7.6.1 Comparison of regression methods

Several standard dimensionality reduction techniques were tested. We used the same methods as the ones presented in Section 5.2.1, namely low p-value features (pval), principal component analysis (PCA), and all features (all). For regression, we used the same regression techniques as the ones presented in Section 5.2.1: ridge regression (ridge), and random forest (RF); ordinary least squares (OLS) was dropped due to its consistently inferior results. Additionally, we used LASSO regression, which, similarly to ridge regression, minimizes the sum of squared errors between the observed and predicted responses of a linear model, but the regularization term multiplies the l_1-norm of the regression coefficients (instead of the l_2-norm in ridge regression), resulting in sparse coefficients and preventing the model to over-fit [131].

Table 7.10 displays the inference results for personality and hirability scores obtained using different dimensionality reduction and regression techniques. For personality variables, only extraversion and openness to experience were inferred significantly more accurately than the baseline-average model, with R^2 values up to 0.27 for extraversion and 0.20 for openness to experience. Under the light of the analyses performed in Sections 7.2.5 and 7.5, it is not surprising that these variables were the only ones accurately predicted: extraversion and openness to experience were the personality variables with the highest inter-rater agreements and with the largest number of correlated nonverbal cues. To contextualize these results,

Biel *et al.* obtained inference results of $R^2 = .36$ for extraversion and $R^2 = .10$ for openness to experience in a dataset of video blogs [29]. This suggests that although video blogs seem to be a better-suited setting to predict extraversion, this task can also be achieved in video resumes, while more accurate results were achieved for openness to experience.

For hirability, all variables could be inferred significantly more accurately than the baseline-average model, with R^2 values up to 0.19 for overall impression, 0.15 for overall hirability, 0.12 for professional skills, 0.21 for social skills, and 0.20 for communication skills. Social and communication skills were the hirability variables with the most accurate prediction results, and were also the variables with the largest number of correlated cues (see Section 7.5). In comparison with the results obtained for the inference of hirability in employment interviews (Chapter 5), R^2 values were lower, but the results were more significant due to the increase of data points by over an order of magnitude, showcasing the benefits of using large datasets.

In terms of regression methods, random forest with no dimensionality reduction consistently yielded the most accurate results. Dimensionality reduction methods did not increase the prediction accuracy, and this can be explained by the low feature dimensionality of the original space ($D = 92$) compared to the number of data points. The results obtained here differ from the ones obtained on the employment interview dataset, where ridge regression yielded more accurate results than random forest (see Section 5.2), suggesting that random forest requires a larger number of data points than ridge regression to be accurate.

Independently of the inference method and the analyzed social variable, the results obtained here show the feasibility of automatically inferring social impressions to some degree. Moreover, our initial hypothesis stating that nonverbal behavior can be used for the inference of personality and hirability variables holds. Despite the high variance in quality in the video resume dataset and the noise present in the features due to the error-prone nonverbal cue extraction step, positive inference results could be achieved. Furthermore, some of the results can be found in concordance with recent literature on nonverbal analysis of social videos [29].

7.6.2 Feature group analysis

In order to understand the basis on which personality and hirability impressions were formed beyond the correlation analysis performed in Section 7.5, we conducted a feature group analysis. Six feature groups were defined, based on the type of nonverbal behavior they were designed to represent, namely vocal, looking turns, proximity, head motion, overall motion (WMEI), and video statistics. To assess their predictive validity, each feature group was used individually in a regression task. Because random forest with no dimensionality reduction was consistently observed to be the most accurate method for the inference of personality and hirability impressions (see Section 7.6.1), this regression method was used here.

Table 7.11 displays the performance[1] (R^2 and $RMSE$) for the inference of personality and

[1] Note that due to the random nature of random forest, the inference results may slightly differ between

Table 7.10 – Performance (R^2 and $RMSE$) for the inference of personality and hirability impressions using different dimensionality reduction and regression methods ($^*p < 10^{-3}$, $^\dagger p < 10^{-4}$ for $RMSE$, and $p > 10^{-3}$ for values with no symbols). The best achieved result for each variable is highlighted in bold. $N = 882$.

	Extra.		**Agree.**		**Consc.**		**Neuro.**		**Open.**	
	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$
Baseline-Avg	0.00	0.61	0.00	0.39	0.00	0.45	0.00	0.39	0.00	0.51
All-Lasso	0.23	0.54†	0.01	0.39	0.00	0.45	-0.06	0.40	0.14	0.48†
All-Ridge	0.24	0.53†	0.03	0.38	0.02	0.45	0.00	0.39	0.17	0.47†
All-RF	**0.27**	**0.52**†	0.06	0.38	0.03	0.44	0.00	0.39	**0.20**	**0.46**†
Pval-Ridge	0.23	0.53†	0.04	0.38	0.00	0.45	-0.01	0.39	0.18	0.47†
Pval-Lasso	0.22	0.54†	0.04	0.38	-0.01	0.46	-0.03	0.40	0.17	0.47†
Pval-RF	0.25	0.53†	0.04	0.38	0.00	0.45	-0.06	0.40	0.18	0.46†
PCA-Ridge	0.23	0.53†	0.03	0.38	0.02	0.45	0.00	0.39	0.17	0.47†
PCA-Lasso	0.22	0.54†	-0.01	0.39	-0.00	0.45	-0.05	0.40	0.13	0.48*
PCA-RF	0.23	0.54†	0.04	0.38	0.02	0.45	0.01	0.39	0.17	0.47†

	Ov. Impression		**Ov. Hirability**		**Pro. Skills**		**Social Skills**		**Comm. Skills**	
	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$
Baseline-Avg	0.00	0.62	0.00	0.62	0.00	0.59	0.00	0.63	0.00	0.70
All-Lasso	0.12	0.58	0.09	0.59	0.08	0.56	0.11	0.59	0.09	0.66
All-Ridge	0.11	0.59*	0.09	0.60*	0.09	0.56*	0.12	0.59†	0.14	0.65†
All-RF	**0.18**	**0.56**†	**0.15**	**0.57**†	**0.12**	**0.55**†	**0.21**	**0.56**†	**0.20**	**0.62**†
Pval-Ridge	0.13	0.58†	0.09	0.60*	0.09	0.56*	0.12	0.59†	0.13	0.65†
Pval-Lasso	0.14	0.57†	0.11	0.59*	0.10	0.56	0.09	0.60	0.08	0.67
Pval-RF	0.17	0.57†	0.14	0.58†	0.11	0.56†	0.20	0.56†	0.18	0.63†
PCA-Ridge	0.12	0.58†	0.09	0.60*	0.09	0.56*	0.12	0.59†	0.14	0.65†
PCA-Lasso	0.10	0.59	0.08	0.60	0.05	0.57	0.08	0.60	0.08	0.67
PCA-RF	0.14	0.57†	0.10	0.59†	0.10	0.56†	0.13	0.59†	0.15	0.64†

hirability scores using different feature groups and random forest with no dimensionality reduction, with D denoting the number of cues in each feature group. One can observe that despite the low number of vocal cues correlated with personality and hirability variables (see Section 7.5), the vocal feature group showed the highest predictive validity for all hirability variables and for extraversion and openness to experience. In particular, vocal cues alone achieved to explain 17% of the variance of extraversion and were marginal for the other variables. Similar results were also obtained on video blogs, with vocal cues explaining 31% of the variance of the extraversion trait [29]. The relatively large ($D = 33$) number of cues contained in the vocal feature group fails to fully explain these results: head motion ($D = 24$) had a larger number of cues correlated with most variables, but showed poor predictive validity. To understand these results, we performed principal component analysis (PCA) on the feature groups and counted the number of components needed to explain 90% of the variance. For head motion, only 5 components achieved to explain over 90% of the variance, while 11 were necessary for vocal cues, showcasing the greater variety of the vocal feature group. We believe that this aspect at least partially explains the high predictive validity of this feature group. In addition to vocal cues, overall motion measured using weighted motion energy images, as well as basic video statistics showed marginally good predictive validity (except for extraversion where it was more significant).

Tables 7.10 and 7.11.

Table 7.11 – Performance (R^2 and $RMSE$) for the inference of personality and hirability scores using different feature groups and random forest with no dimensionality reduction ($^{*}p < 10^{-3}$, $^{\dagger}p < 10^{-4}$ for $RMSE$, and $p > 10^{-3}$ for values with no symbols). The best achieved result for each variable is highlighted in bold. $N = 882$.

		Extra.		**Agree.**		**Consc.**		**Neuro.**		**Open.**	
	D	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$
Baseline-Avg	-	0.00	0.61	0.00	0.39	0.00	0.45	0.00	0.39	0.00	0.51
Vocal	33	**0.17**	**0.56**†	0.02	0.39	0.03	0.44	-0.02	0.40	**0.10**	**0.49**†
Looking Turns	9	0.02	0.60	-0.10	0.41	-0.17	0.49	-0.19	0.43	-0.01	0.52
Proximity	8	0.11	0.58*	-0.09	0.41	-0.08	0.47	-0.10	0.41	0.06	0.50
Head Motion	24	0.08	0.59	-0.05	0.40	-0.07	0.47	-0.06	0.40	0.08	0.49
Ov. motion	10	0.14	0.56†	-0.02	0.40	-0.02	0.46	-0.04	0.40	0.08	0.49
Video Stats	8	0.11	0.58†	-0.04	0.40	-0.04	0.46	-0.04	0.40	0.08	0.49
All-NVB	92	0.28	0.52†	0.05	0.38	0.03	0.44	0.00	0.39	0.20	0.46†
		Ov. Impression		**Ov. Hirability**		**Pro. Skills**		**Social Skills**		**Comm. Skills**	
	D	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$
Baseline-Avg	-	0.00	0.62	0.00	0.62	0.00	0.59	0.00	0.63	0.00	0.70
Vocal	33	**0.10**	**0.59**†	**0.07**	**0.60**	**0.06**	**0.57**	**0.09**	**0.60***	**0.13**	**0.65**†
Looking Turns	9	-0.08	0.65	-0.08	0.65	-0.06	0.61	-0.03	0.64	-0.06	0.72
Proximity	8	-0.00	0.62	-0.03	0.63	-0.07	0.61	0.01	0.62	-0.02	0.70
Head Motion	24	0.03	0.61	0.02	0.62	-0.01	0.59	0.06	0.61	0.04	0.68
Ov. motion	10	0.04	0.61	0.03	0.61	0.03	0.58	0.09	0.60	0.04	0.68
Video stats	8	0.07	0.60	0.06	0.60	0.02	0.58	0.08	0.60	0.07	0.67
All-NVB	92	0.17	0.56†	0.15	0.57†	0.11	0.56†	0.20	0.56†	0.20	0.62†

Overall, for personality and hirability variables, using all nonverbal cues yielded the best inference results. In other words, combining feature groups strongly improved the prediction accuracy. One hypothesis to interpret this observation is that each feature group explained a different part of the variance in the data, which added up when put in combination.

7.7 Conclusion

In this chapter, we analyzed the formation of personality and hirability impressions on conversational video resumes. To the best of our knowledge, this work constitutes the first computational study on video resumes; it is also the first investigating video resumes from a nonverbal standpoint.

As a first step, we collected a dataset of 939 conversational English-speaking video resumes hosted on YouTube. Crowdsourced annotations of basic facts, demographics, perceived skills, and first impressions of hirability and personality were collected using Amazon Mechanical Turk, and most variables were found to be reliable upon analysis of the interrater agreement, not only suggesting that raters completed the HITs conscientiously, but also that the annotations of first impressions of job-related skills, hirability, and the personality variables of extraversion and openness to experience by unacquainted naïve judges was a feasible task. This in itself is a positive result regarding the use of crowdourcing to obtain impressions at large scale. The personality variables of agreeableness, conscientiousness, and neuroticism however had low interrater agreement, suggesting that these traits were more difficult to rate in this setting. The analysis of the demographics showed that the job-seekers composing the

video resume dataset were mainly young and were applying for internship or junior positions. Demographics varied across job categories: the professional category (engineers, computer scientists, and scientists) was mainly composed of young Indian males, whereas hospitality was dominated by Caucasians, with an equal number of men and women. This is probably biased by the geographic origin of the posts, which YouTube does not give access to.

To understand the structure underlying the perceived skills, we conducted a data-driven clustering analysis and results showed that skills could be grouped into three high-level clusters, namely professional skills, communication skills, and social skills. Nonverbal cues were automatically extracted from the visual and audio modalities to obtain a feature representation of the video resumes. As a first step, we performed a correlation analysis between nonverbal cues and the organizational social constructs of personality and hirability, and found that head motion, looking turns, proximity, and overall motion were correlated with extraversion, openness to experience, social skills, and communication skills, whereas only proximity and head motion cues were associated to overall hirability and overall impression. As a second step, we evaluated several regression methods for the inference of personality and hirability. Results demonstrated the feasibility of automatically inferring hirability variables as well as extraversion and openness to experience in video resumes, achieving R^2 results up to 27% in a proper cross-validation experimental setting, thus confirming our initial hypothesis that first impressions on video resumes were at least partly based on nonverbal behavior.

Several possible research directions can be considered for future work. First, the accuracy of the nonverbal cue extraction process could be improved despite the challenging nature of the video resumes. Second, more behavioral cues could be extracted, such as facial expressions, true gaze, head nods, or verbal content; moreover, non-behavioral cues could also be interesting to extract and analyze, such as the presence of music or changes of shots. Also, extracting multimodal cues previously documented in psychology such as speaking while looking could also improve the prediction accuracy. Third, other inference tasks could be performed, such as the classification of job categories, or the inference of hirability and personality conditioned on the job categories. Last, all analyses in this chapter were conducted based on crowdsourced annotations of social variables by naïve raters. We strongly believe that a comparison with expert raters (*e.g.*, human resources professionals) could be beneficial for a deeper understanding of how first impressions are made in organizational settings; such gold-standard annotations would allow us to assess and compare the predictive validity of automatically nonverbal cues and crowdsourced annotations.

8 Conclusion

In this dissertation, we proposed and evaluated a computational framework for the inference of organizational social constructs in face-to-face employment interviews and online conversational video resumes. We addressed this problem from a behavioral perspective where sensing, cue extraction, and inference were automated. To our knowledge, the analysis of hirability from a computational standpoint had remained unexplored before this work. To address this issue, we first designed and collected a dataset of 62 real employment interviews, where applicants were applying for a paid marketing job. To obtain a detailed feature representation of the interviews, we extracted behavioral cues from the visual and audio modalities for both the job applicant and the interviewer. We then conducted an inference analysis and achieved to explain up to 36% of the variance, demonstrating that automatically predicting hirability was a promising task. We then analyzed the predictive validity of thin slices, and showed that although nonverbal behavior extracted from thin slices was not as predictive as the full interview, it still achieved good performance in the prediction of the interview outcome. Last, we collected and analyzed a set of 939 online conversational video resumes, and we achieved to explain up to 27% of the variance of organizational social constructs, using video crowdsourcing to collect impressions of a rich number of variables and simple automatically extracted nonverbal cues.

The rest of this chapter is structured as follows. In Section 8.1, we summarize the main contributions of this thesis chapter by chapter. In Section 8.2, we point out the limitations of our work and discuss possible avenues for future work.

8.1 Main contributions

In Chapter 2, we contextualized our work by reviewing the literature in relation to the computational analysis of employment interviews and online conversational video resumes. Due to the multidisciplinary nature of this thesis, the related work spanned several research fields, such as nonverbal communication, organizational psychology, audio processing, computer vision, and social computing.

In Chapter 3, we presented the collection of the SONVB employment interview dataset, which served as a basis for the main part of this thesis. This data collection was motivated by the lack of a publicly available dataset of job interviews. The dataset comprises 62 real job interviews, where participants were applying for a marketing job. Both the applicant and the interviewer were recorded using multiple modalities (audio, video, and Kinect). Psychometric questionnaires were completed by job applicants to assess their personality, intelligence, and communication skills. Annotations of hirability impressions were collected using three different schemes, depending on the level of expertise of the raters (master's students in organizational psychology and human resource professionals) and the material available (audio-visual recordings and questionnaire data output).

In Chapter 4, we presented our method to extract behavioral cues to obtain a feature representation of the employment interviews. We first developed a multimodal method to extract head nods in natural interactions, and results showed that using the audio self-context improved the detection of head nods. Then, three types of behavioral cues were extracted. First, we extracted audio-visual nonverbal cues for both the interviewer and the applicant. To select what behavioral features to extract, we searched the psychology literature for cues consistently reported to play a role in job interviews, and used available computational methods to extract behavioral features. Second, we extracted body communication cues to understand the role of an applicant's hand gestures and seated posture. Because of the difficulty to track body posture in seated configurations and to benchmark the use of posture and gestures in an ideal case, we used manual annotations of body communication. Additionally, we automatically extracted descriptors based on hand velocity as a raw representation of hand gestures. Third, we collected manual speech transcriptions and extracted linguistic and paralinguistic features using the Linguistic Inquiry Word Count (LIWC) to examine the role of verbal content in the formation of hirability impressions. LIWC is a word categorization system that relates linguistic and paralinguistic categories to psychological constructs, and each word from the transcript found in the dictionary increments one or several categories to which the word belongs.

In Chapter 5, we proposed a computational framework for the automatic prediction of hirability in real job interviews, using applicant and interviewer behavioral cues extracted from the audio and visual modalities. First, we performed a correlation analysis between nonverbal cues and hirability impressions and found that not only applicant cues were correlated with the hirability scores, but interviewer cues, too. Second, we evaluated several prediction methods for a regression task. Results demonstrated the feasibility of predicting hirability scores based on automatically extracted nonverbal cues, and validated our proposed framework, with R^2 values of up to 36%. Third, we analyzed the predictive validity of feature groups and observed that the most predictive groups were the applicant audio cues and the interviewer visual cues. This finding suggests that the interviewer produced behavioral responses which were conditioned on the quality of the job applicant by displaying more visual back-channels. Fourth, we analyzed the use of psychometric questionnaires widely used in the personnel selection process for the prediction of hirability scores: results showed that questionnaires

were not predictive of hirability and did not improve the inference results when combined with nonverbal cues. Fifth, we investigated the use of a mixture of manual and automatic applicant body cues for the prediction of hirability and personality. The results showed that the prediction of these variables using body cues was a promising task, and that including the speaking status in the extraction process improved the prediction accuracy. Last, we analyzed the predictive validity of verbal content, represented by LIWC-based linguistic and paralinguistic features from manual transcriptions. Verbal content by itself was not able to yield accurate prediction results, and was unable to improve the prediction accuracy when combined to nonverbal behavior, suggesting that raters might have formed their hirability impressions based on nonverbal behavior rather than verbal content.

In Chapter 6, we analyzed thin slices of job interviews, where slices were defined by the specific questions of the interview structure. To assess the observer predictive validity of thin slices, annotations were completed on snippets of videos defined by the interview question/answer slices, and correlation analyses showed that predicting hirability from automatically extracted nonverbal cues from the full interview yielded results similar to using annotations obtained by human resource professionals based on thin slices. Audiovisual nonverbal cues were extracted for both the applicant and the interviewer for each slice, and applicant cues related to voice characteristics, speaking turns, and head motion were consistently correlated with the hiring decision, with the exception of applicant speaking time which was positively correlated for the full interview, but negatively correlated for some slices. Interviewer pitch standard deviation and short utterances were also consistently correlated with the full interview hiring decision score, but the interviewer's nodding behavior was observed to be conflicting: interviewer nods taken from the full interaction were positively correlated, while the valence was reversed in thin slices. We then performed a regression task, where the goal was to infer the full interview ratings based on the nonverbal cues extracted from the full slices. Although behavioral cues extracted from thin slices were found not to be as accurate as the full interaction, they were still predictive of the interview outcome: the best results obtained from thin slices were competitive compared to the observer predictive validity, with R^2 of up to 0.34. No slice stood out in terms of predictive validity, suggesting that the observed nonverbal behavior did not drastically change from one slice to another. To understand the basis on which raters formed their hirability impressions, we examined the accuracy stemming from person- and modality-based feature groups, and applicant audio features were observed to yield the most accurate results. To examine the predictive validity of interview questions, we split the slices into three thin slice cases, question-only, answer-only, and whole-slice. Questions taken alone were found to consistently yield negative results, whereas answers predicted the hiring decision score to a lesser degree than using the full interview; moreover, adding the questions to the answers (*i.e.* using the whole-slice case) did not significantly improve the predictive validity, which suggests that raters made their impression based on the applicant's speaking behavior.

In Chapter 7, we analyzed the formation of personality and hirability impressions in conversational video resumes. To the best of our knowledge, this work constitutes the first computational study on video resumes; it is also the first investigating video resumes from a nonverbal

standpoint. As a first step, we collected a dataset of 939 conversational English-speaking video resumes hosted on YouTube, which represents over one order of magnitude more data than what we collecte din the lab. Crowdsourced annotations of demographics, perceived skills, and first impressions of hirability and personality were collected using Amazon Mechanical Turk, and most variables were found to be reliable upon analysis of the interrater agreement, not only suggesting that raters competed the HITs conscientiously, but also that the annotations of first impressions of job-related skills, hirability, and the personality variables of extraversion and openness to experience by unacquainted naïve judges was a feasible task in a crowdsourcing setting. The personality variables of agreeableness, conscientiousness, and neuroticism however had low interrater agreement, suggesting that these traits were more difficult to rate in this video genre. The analysis of the demographics showed that the job-seekers composing the video resume dataset were mainly young and were applying for internship or junior positions. Demographics varied across job categories: the professional category (engineers, computer scientists, and scientists) was mainly composed of young Indian males, whereas hospitality was dominated by Caucasians, with an equal number of men and women. To understand the structure underlying the perceived skills, we conducted a data-driven clustering analysis and results showed that skills could be grouped into three high-level clusters, namely professional skills, communication skills, and social skills. Nonverbal cues were automatically extracted from the visual and audio modalities to obtain a feature representation of the video resumes. As a first step, we performed a correlation analysis between nonverbal cues and the organizational social constructs of personality and hirability, and found that head motion, looking turns, proximity and overall motion were correlated with extraversion, openness to experience, social skills, and communication skills, whereas only proximity and head motion cues were associated to overall hirability and overall impression. As a second step, we evaluated several regression methods for automatic inference of personality and hirability. Results demonstrated the feasibility of automatically inferring hirability variables as well as extraversion and openness to experience in video resumes, achieving R^2 results up to 27%, thus confirming our initial hypothesis that first impressions on video resumes were at least partly based on nonverbal behavior.

8.2 Limitations and future work

While our work on the computational analysis of employment interviews contributed several findings to the literature in organizational psychology and social computing, it also has some limitations. First, the size of the SONVB interview dataset is relatively small ($N = 62$) despite our best efforts to collect data. The collection of real employment interviews for academic research is a challenging task, as it requires a significant amount of upstream work before the applicant can be recorded in the lab, such as advertising the study, contacting the participant, scheduling the interview, etc. As a consequence of the relatively small size of the dataset, the significance levels of some results was limited. Furthermore, elaborate inference methods involving a large number of parameters could not be used due to over-fitting issues.

Second, although a broad array of behavioral cues were extracted and examined in detail, important types of nonverbal behavior were not examined in this dissertation. For instance, facial expressions are strongly related to affective states, which might influence how applicants are perceived during employment interviews. The high resolution of the video recordings would likely allow the accurate extraction of facial expressions, allowing us to investigate their relationship with first impressions in job interview. One anticipated challenge of using facial expressions is that they will be affected by speech, which is a result observed in recent work [30]. Another important visual cue that was not analyzed in this dissertation is gaze. Gaze is an important nonverbal behavior because it is used to regulate the flow of communication, monitor feedback, reflect cognitive activity, and communicate the nature of the interpersonal relationship [78], and was found to be correlated with hiring decisions [68]. The RGB-D data recorded from the interviews would allow us to apply state-of-the-art gaze estimation methods [54], opening the possibility to investigate the role of gaze in the formation of first impressions in employment interviews.

A third limitation of this work is related to the generalization of the results. In the job interview dataset, job applicants were applying for a simple marketing job, therefore one could argue that the results obtained in this dissertation are limited to this type of jobs. We believe in the necessity to examine the role of nonverbal behavior in the formation of hirability impressions for other types of occupations. We believe that job categories requiring interpersonal and communication skills are the best suited for the development of automated methods to infer hirability impressions and rank job applicants. Jobs related to sales or hospitality constitute good examples of such types of occupations. Clearly, this hypothesis would have to be validated in the future.

Fourth, in the conversational video resume dataset, one limitation of our work is nonverbal cue validity, *i.e.* verifying that the extracted nonverbal cues are appropriately capturing the conversational aspect that they are supposed to measure. Given the relatively large number of video resumes in the dataset, assessing cue validity is a challenging task. Related to this aspect, the accuracy of the nonverbal cue extraction process could be improved despite the challenging nature of the dataset.

Fifth, more behavioral cues could be extracted from the conversational video resume dataset, such as facial expressions, gaze, head nods, or verbal content; moreover, non-behavioral cues could also be interesting to extract and analyze, such as the presence of music or changes of shots. Extracting multimodal cues such as speaking while looking could also improve the prediction accuracy. Also, other inference tasks could be performed, such as the classification of job categories, or the inference of hirability and personality conditioned on the job categories. There are also tasks for which we have ground-truth data generated by the crowdsourcing experiments.

Sixth, all analyses of the conversational video resume dataset were conducted based on crowdsourced annotations of social variables by naïve raters. We believe that a comparison with

expert raters (*e.g.*, human resources professionals) could be beneficial to gain deeper understanding of how first impressions are made in organizational settings; such gold-standard annotations would allow us to assess and compare the predictive validity of automatically nonverbal cues and the crowdsourced annotations themselves.

Seventh, the models trained to infer hirability impressions from behavioral cues were based on manual annotations of hirability obtained from human resource professionals, psychology researchers, or naïve judges. These supervised models learn the mappings between behavioral features and hirability variables, therefore aim at mimicking the impressions of raters. One limitation of this approach is that models implicitly learn the possible biases of raters, which might not be desirable for the development of a neutral tool to rate applicants in job interviews or video resumes. We believe that Amazon Mechanical Turk constitutes an interesting tool to address this issue as a large number of annotations could be collected for each applicant. Analyzing the distribution of hirability impressions would allow not only to smooth out the biases by aggregating all annotations, but also to study the biases of the raters themselves, which could be associated with stereotypes.

Finally, while accurately predicting hirability is in our opinion relevant in and of itself, it does not tell whether the right choice was made; job performance is related to hirability, but is a different social dimension. Accurately predicting the most performing applicants using an automated method could therefore be seen as the ultimate task, but this is by definition a difficult problem as many different factors which are not necessarily elicited in a job interview can affect job performance; some might even be impossible to sense. As a last point, the validity of employment interviews for selecting the most performant candidates is still an open question in organizational psychology [110], which suggests that other settings should also be considered.

Bibliography

[1] Amazon Mechanical Turk. Available: https://www.mturk.com [online].

[2] American Time Use Survey. Available: http://www.bls.gov/tus/tables/a4_1112.pdf [online].

[3] FFmpeg: A complete, cross-platform solution to record, convert and stream audio and video. Available: https://www.ffmpeg.org/ [online].

[4] Hireiq. Available: http://www.hireiqinc.com/ [online].

[5] Hirevue. Available: http://www.hirevue.com/ [online].

[6] Microcone: Intelligent microphone array for groups. Available: http://www.dev-audio.com/products/microcone/ [online].

[7] MIT speech feature extraction code. Available: http://groupmedia.media.mit.edu/data.php [online].

[8] OpenKinect/Libfreenect: Drivers and libraries for the Xbox Kinect device on Windows, Linux, and OS X. Available: http://openkinect.org/ [online].

[9] Praat: Doing phonetics by computer. Available: http://www.fon.hum.uva.nl/praat/ [online].

[10] Sparkhire. Available: https://www.sparkhire.com/ [online].

[11] Video recruit. Available: https://www.video-recruit.com/ [online].

[12] Wavesurfer. Available: http://sourceforge.net/projects/wavesurfer/ [online].

[13] Wonderlic cognitive ability test. Available: http://www.wonderlic.com/assessments/ability/cognitive-ability-tests/classic-cognitive-ability-test [online].

[14] Youtube data API (v3). Available: https://developers.google.com/youtube/v3/ [online].

[15] Youtube-dl: Download videos from youtube (and more sites). Available: http://rg3.github.io/youtube-dl/ [online].

[16] J. Allwood and L. Cerrato. A study of gestural feedback expressions. In *Proceedings of the First Nordic Symposium on Multimodal Communication*, 2003.

[17] N. Ambady, F. J. Bernieri, and J. A. Richeson. Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology*, 32:201–271, 2000.

[18] N. Ambady, M. Hallahan, and R. Rosenthal. On judging and being judged accurately in zero-acquaintance situations. *Journal of Personality and Social Psychology*, 69(3):518–529, 1995.

[19] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2):256–274, 1992.

[20] N. Anderson and V. Shackleton. Decision making in the graduate selection interview: A field study. *Journal of Occupational Psychology*, 63(1):63–76, 1990.

[21] X. Anguera, C. Wooters, and J. Hernando. Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech and Language Processing*, 15(7):2011–2023, 2007.

[22] M. R. Barrick and M. K. Mount. The Big Five personality dimensions and job performance: A meta-analysis. *Journal of Personnel Psychology*, 44(1):1–26, 1991.

[23] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: Machine learning and application to spontaneous behavior. In *Proceeding of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[24] S. Basu. *Conversational scene analysis*. PhD thesis, Massachussetts Institute of Technology, 2002.

[25] L. M. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe. Please, tell me about yourself: automatic personality assessment using short self-presentations. *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2011.

[26] T. Baur, I. Damian, P. Gebhard, K. Porayska-Pomsta, and E. Andre. A job interview simulation: Social cue-based interaction with a virtual character. In *Proceedings of the IEEE/ASE International Conference on Social Computing (SocialCom)*, 2013.

[27] J.-I. Biel, O. Aran, and D. Gatica-Perez. You are known by how you vlog: Personality impressions and nonverbal behavior in YouTube. In *Proceedings of the AAAI International Conference on Web and Social Media (ICSWM)*, 2011.

[28] J.-I. Biel and D. Gatica-Perez. VlogSense: Conversational behavior and social attention in YouTube. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2(3):1–20, 2010.

[29] J.-I. Biel and D. Gatica-Perez. The YouTube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1):41–55, 2013.

[30] J.-I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez. FaceTube: Predicting personality from facial expressions of emotion in online conversational video. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2012.

[31] J.-I. Biel, V. Tsiminaki, J. Dines, and D. Gatica-Perez. Hi YouTube! Personality impressions and verbal content in social video. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2013.

[32] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.

[33] J. Bjerke. 63 percent of companies use video interviews. Available: https://www.recruiter.com/i/63-percent-of-companies-use-video-interviews/ [online], 2013.

[34] L. Breiman. Random forests. *Machine Learning*, 45(1):1–35, 2001.

[35] D. F. Caldwell and J. M. Burger. Personality characteristics of job applicants and success in screening interviews. *Journal of Personnel Psychology*, 51(1):119–136, 1998.

[36] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski. Workshop on computational personality recognition: Shared task. In *Proceedings of the Workshop on Computational Personality Prediction*, 2013.

[37] L. Chen, M. Martin, and M. Ma. An initial analysis of structured video interviews by using multimodal emotion detection. In *Proceedings of the International Workshop on Emotion Representations and Modelling for Human-Computer Interaction Systems (ERM4HCI)*, 2014.

[38] C. Chung and J. Pennebaker. The psychological functions of function words. In *Social Communication*, chapter 12, pages 343–359. New York: Psychology Press, 2007.

[39] J. F. Cohn, Z. Ambadar, and P. Ekman. Observer-based measurement of facial expression with the facial action coding system. In *The handbook of emotion elicitation and assessment*, pages 203–221. 2005.

[40] J. F. Cohn and K. L. Schmidt. The timing of facial motion in posed and spontaneous smiles. *Journal of Wavelets, Multi-resolution & Information Processing*, 2(2):1–12, 2004.

[41] M. S. Cole, H. S. Feild, W. F. Giles, and S. G. Harris. Recruiters' inferences of applicant personality based on resume screening: Do paper people have a personality? *Journal of Business and Psychology*, 24(1):5–18, 2008.

[42] D. Collins. The 500-Year Evolution Of The Resume. Available: http://www.businessinsider.com/how-resumes-have-evolved-since-their-first-creation-in-1482-2011-2?op=1 [online], 2011.

[43] P. T. Costa and R. R. McCrae. *Neo PI-R professional manual*. Psychological Assessment Resources, 1992.

[44] J. Curhan and A. Pentland. Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology*, 92(3):802–811, 2007.

[45] T. DeGroot and J. Gooty. Can nonverbal cues be used to make meaningful personality attributions in employment interviews? *Journal of Business and Psychology*, 24(2):179–192, 2009.

[46] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, 2012.

[47] L. Dong, Y. Jin, L. Tao, and G. Xu. Recognition of multi-pose head gestures in human conversations. In *Proceedings of the International Conference on Image and Graphics (ICIG)*, 2007.

[48] A. Doyle. Skills list for resumes. Available: http://jobsearch.about.com/od/list/fl/list-of-skills-resume.htm [online].

[49] E. Faliagka, K. Ramantas, A. Tsakalidis, and G. Tzimas. Application of machine learning algorithms to an online recruitment system. In *Proceedings of the International Conference on Internet and Web Applications and Services (ICIW)*, 2012.

[50] S. Favre, H. Salamin, J. Dines, and A. Vinciarelli. Role recognition in multiparty recordings using social affiliation networks and discrete distributions. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, 2008.

[51] S. Feese, B. Arnrich, G. Tröster, B. Meyer, and K. Jonas. Detecting posture mirroring in social interactions with wearable sensors. In *Proceedings of the IEEE International Symposium on Wearable Computers (ISWC)*, 2011.

[52] J. L. Fleiss, B. Levin, and M. C. Paik. The measurement of interrater agreement. In *Statistical Methods for Rates and Proportions*, pages 598–626. John Wiley and Sons, 3rd edition, 2003.

[53] R. J. Forbes and P. R. Jackson. Non-verbal behaviour and the outcome of selection interviews. *Journal of Occupational Psychology*, 53(1):65–72, 1980.

[54] K. A. Funes Mora and J.-M. Odobez. Person independent 3D gaze estimation from remote RGB-D cameras. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2013.

[55] N. Gada-Jain. Intentional synchrony effects on job interview evaluation. Master's thesis, University of Toledo, 1999.

[56] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*, 27(12):1775–1787, 2009.

[57] R. Gifford, C. F. Ng, and M. Wilkinson. Nonverbal cues in the employment interview: Links between applicant qualities and interviewer judgments. *Journal of Applied Psychology*, 70(4):729–736, 1985.

[58] S. Gorga and K. Otsuka. Conversation scene analysis based on dynamic Bayesian network and image-based gaze detection. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2010.

[59] S. Gosling. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6):504–528, 2003.

[60] S. Gosling, S. Gaddis, and S. Vazire. Personality impressions based on Facebook profiles. In *Proceedings of the AAAI International Conference on Web and Social Media (ICSWM)*, 2007.

[61] S. D. Gosling, S. J. Ko, T. Mannarelli, and M. E. Morris. A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology*, 82(3):379–398, 2002.

[62] U. Hadar, T. J. Steiner, and F. Clifford Rose. Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4):214–228, 1985.

[63] U. Hadar, T. J. Steiner, E. C. Grant, and F. Clifford Rose. Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2(1):35–46, 1983.

[64] D. W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):478–500, 2010.

[65] A. M. F. Hiemstra. *Fairness in Paper and Video Resume Screening*. PhD thesis, Erasmus University Rotterdam, Netherlands, 2013.

[66] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Journal of Technometrics*, 12(1):55–67, 1970.

[67] M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard. MACH: My Automated Conversation coacH. In *Proceedings of the ACM International Conference on Ubiquitous Computing (Ubicomp)*, 2013.

[68] J. L. Howard and G. R. Ferris. The employment interview context: Social and situational influences on interviewer decisions. *Journal of Applied Social Psychology*, 26(2):112–136, 1996.

[69] A. I. Huffcut, J. M. Conway, P. L. Roth, and N. J. Stone. Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, 86(5):897–913, 2001.

[70] F. Iacobelli, A. J. Gill, S. Nowson, and J. Oberlander. Large scale personality classification of bloggers. In *Proceedings of the IEEE International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2011.

[71] A. S. Imada and M. D. Hakel. Influence of nonverbal communication and rater proximity on impressions and decisions in simulated employment interviews. *Journal of Applied Psychology*, 62(3):295–300, 1977.

[72] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):501–513, 2009.

[73] D. B. Jayagopi, D. Sanchez-Cortes, K. Otsuka, J. Yamato, and D. Gatica-Perez. Linking speaking and looking behavior patterns with group composition, perception, and performance. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2012.

[74] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 2nd edition, 2002.

[75] A. Kapoor and R. W. Picard. A real-time head nod and shake detector. In *Proceedings Workshop on Perceptive User Interfaces (PUI)*, 2001.

[76] J. F. Kelly and E. H. O'Brien. Using video resumes to teach deaf college students job search skills and improve their communication. *American Annals of the Deaf*, 137(5):404–410, 1992.

[77] K. J. Kemp, L. M. Bobbitt, M. B. Beauchamp, and E. A. Peyton. Using one-minute video résumés as a screening tool for sales applicants. *Journal of Marketing Development and Competitiveness*, 7(1):84–92, 2013.

[78] M. L. Knapp and J. A. Hall. *Nonverbal communication in human interaction*. Wadsworth, Cengage Learning, 7th edition, 2009.

[79] M. Kotti, V. Moschou, and C. Kotropoulos. Speaker segmentation and clustering. *Signal Processing*, 88(5):1091–1124, 2007.

[80] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, (33):159–174, 1977.

[81] B. Lepri, N. Mana, A. Cappelletti, and F. Pianesi. Automatic prediction of individual performance from "thin slices" of social behavior. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2009.

[82] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe. Connecting meeting behavior with extraversion - A systematic study. *IEEE Transactions on Affective Computing*, 3(4):443–455, 2012.

[83] R. C. Liden, C. L. Martin, and C. K. Parsons. Interviewer and applicant behaviors in employment interviews. *Academy of Management Journal*, 36(2):372–386, 1993.

[84] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The Computer Expression Recognition Toolbox (CERT). In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2011.

[85] F. Lu, Y. Sugano, O. Takahiro, and Y. Sato. Inferring human gaze from appearance via adaptive linear regression. In *Proceedings of the IEEE international Conference on Computer Vision (ICCV)*, 2011.

[86] A. Madan. *Thin Slices of Interest*. PhD thesis, Massachusetts Institute of Technology, 2005.

[87] A. Madan, R. Caneel, and A. Pentland. Voices of attraction, 2004.

[88] A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, L. S. Nguyen, and D. Gatica-Perez. Body communicative cue extraction for conversational analysis. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2013.

[89] M. A. McDaniel, D. L. Whetzel, F. L. Schmidt, and S. D. Maurer. The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, 79(4):599–616, 1994.

[90] T. V. McGovern. The making of a job interviewee: The effect of nonverbal behavior on an interviewer's evaluations during a selection interview. *Dissertation Abstracts International*, 37(9B):4740–4741, 1976.

[91] T. V. McGovern, B. W. Jones, and S. E. Morris. Comparison of professional versus student ratings of job interviewee behavior. *Journal of Counseling Psychology*, 26(2):176–179, 1979.

[92] D. McNeill. So you think gestures are nonverbal? *Psychological Review*, 92(3):350–371, 1985.

[93] T. D. McShane and F. Purdue. Effects of nonverbal cues and verbal first impressions in unstructured and situational interview settings. *Journal of Applied Human Resources Management Research*, 4(2):137–150, 1993.

[94] L.-P. Morency, I. De Kok, and J. Gratch. Context-based recognition during human interactions: Automatic feature selection and encoding dictionary. In *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI)*, 2008.

[95] L.-P. Morency, I. De Kok, and J. Gratch. Predicting listener backchannels: A probabilistic multimodal approach. *Lecture Notes in Computer Science*, 5208:176–190, 2008.

[96] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. Contextual recognition of head gestures. In *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI)*, 2005.

[97] L. S. Nguyen, D. Frauendorfer, M. Schmid Mast, and D. Gatica-Perez. Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Transactions on Multimedia*, 16(4):1018–1031, 2014.

[98] L. S. Nguyen, A. Marcos-Ramiro, M. Marrón Romera, and D. Gatica-Perez. Multimodal analysis of body communication cues in employment interviews. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2013.

[99] L. S. Nguyen, J.-M. Odobez, and D. Gatica-Perez. Using self-context for multimodal detection of head nods in face-to-face interactions. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2012.

[100] R. E. Nisbett and T. Decamp Wilson. The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35(4):250–256, 1977.

[101] S. Nowson and J. Oberlander. Identifying more bloggers: Towards large scale personality classification of personal weblogs. In *Proceedings of the AAAI International Conference on Web and Social Media (ICSWM)*, 2007.

[102] J.-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4):348–365, 1995.

[103] S. Park, J. Gratch, and L. P. Morency. I already know your answer: Using nonverbal behaviors to predict immediate outcomes in a dyadic negotiation. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2012.

[104] C. K. Parsons and R. C. Liden. Interviewer perceptions of applicant qualifications: A multivariate field study of demographic characteristics and nonverbal cues. *Journal of Applied Psychology*, 69(4):557–568, 1984.

[105] J. W. Pennebaker and L. A. King. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312, 1999.

[106] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro. Multimodal recognition of personality traits in social interactions. In *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI)*, 2008.

[107] M. J. Pickering and S. Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190, 2004.

[108] A. Piolat, R. Booth, C. Chung, M. Davids, and J. Pennebaker. La version francaise du dictionnaire pour le LIWC: Modalités de construction et exemples d'utilisation. *Psychologie Francaise*, 56(3):145–159, 2011.

[109] C. Piotrowski and T. Armstrong. Current recruitment and selection practices: A national survey of fortune 1000 firms. *North American Journal of Psychology*, 8(3):489–496, 2006.

[110] R. A. Posthuma, F. P. Morgeson, and M. A. Campion. Beyond employment interview validity: A comprehensive narrative review of recent research and trends over time. *Journal of Personnel Psychology*, 55(1):81, 2002.

[111] F. Ramseyer and W. Tschacher. Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. *Journal of Consulting and Clinical Psychology*, 79(3):284–95, 2011.

[112] K. G. Rasmussen. Nonverbal behavior, verbal behavior, resume credentials, and selection interview outcomes. *Journal of Applied Psychology*, 69(4):551–556, 1984.

[113] S. Raudys. Feature over-selection. *Structural, Syntactic, and Statistical Pattern Recognition Lecture Notes in Computer Science*, 4109:622–631, 2006.

[114] E. Ricci and J.-M. Odobez. Learning large margin likelihoods for realtime head pose tracking. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2009.

[115] R. E. Riggio. Assessment of basic social skills. *Journal of Personality and Social Psychology*, 51(3):649–660, 1986.

[116] J. A. Rolls and M. Strenkowski. Video technology: Resumes of the future. In *World Conference on Cooperative Education*, 1993.

[117] S. Rothmann and E. P. Coetzer. The big five personality dimensions and job performance. *Journal of Industrial Psychology*, 29(1):68–74, 2003.

[118] S. Ruiz-Correa, D. Santani, and D. Gatica-Perez. The young and the city: Crowdsourcing urban awareness in a developing country. In *Proceedings of the International Conference on IoT in Urban Space (Urb-IoT)*, 2014.

[119] D. Sanchez-Cortes, O. Aran, M. Schmid Mast, and D. Gatica-Perez. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia*, 4(3):816–832, 2012.

[120] D. Sanchez-Cortes, J.-I. Biel, S. Kumano, J. Yamato, K. Otsuka, and D. Gatica-Perez. Inferring mood in ubiquitous conversational video. In *Proceedings of the ACM International Conference on Mobile and Ubiquitous Multimedia (MUM)*, 2013.

[121] D. Sanchez-Cortes, P. Motlicek, and D. Gatica-Perez. Assessing the Impact of Language Style on Emergent Leadership Perception from Ubiquitous Audio. In *Proceedings of the ACM International Conference on Mobile and Ubiquitous Multimedia (MUM)*, 2012.

[122] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. S. Rizzo, and L.-P. Morency. Automatic behavior descriptors for psychological disorder analysis. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2013.

[123] G. F. Schmidt. The effect of thin slicing on structured interview decisions. Master's thesis, University of South Florida, 2007.

[124] J. Shotton and A. Fitzgibbon. Real-time human pose recognition in parts from single depth images. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[125] P. E. Shrout and J. L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

[126] C. Smith. By the numbers: 120 amazing LinkedIn statistics. Available: http://expandedramblings.com/index.php/by-the-numbers-a-few-important-linkedin-stats/ [online].

[127] P. Somol, J. Novovicová, and P. Pudil. Efficient Feature Subset Selection and Subset Size Optimization. chapter 4, pages 1–24. 2010.

[128] Y. Song, L.-P. Morency, and R. Davis. Multimodal human behavior analysis: Learning correlation and interaction across modalities. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2012.

[129] W. Tan and G. Rong. A real-time head nod and shake detector using HMMs. *Expert Systems with Applications*, 25(3):461–466, 2003.

[130] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2009.

[131] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

[132] A. Todorov, A. Mandisodza, A. Goren, and C. Hall. Inferences of competence from faces predict election outcomes. *Science*, 308(5728):1623–1626, 2005.

[133] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2011.

[134] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[135] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[136] W. H. Wiesner and S. F. Cronshaw. A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology*, 61(4):275–290, 1988.

[137] B. Wrede and E. Shriberg. Spotting "hot spots" in meetings: Human judgments and prosodic cues. In *Proceedings of Eurospeech*, 2003.

laurentnguyen

PhD candidate

contact

Ch. des Diablerets 8
1012 Lausanne
Switzerland

+41 (79) 318 8189
lnguyen@idiap.ch

key skills

Multi-Disciplinary
Creative
Open-Minded
Organized
Fast-Learner

languages

French: mother tongue
English: C2
German: B1
Spanish/Italian: A1

programming

Matlab,
Python, Java,
LabView, R, C++

education

2011–2015 **PhD** in Electrical Engineering — EPFL, Switzerland
Computational analysis of nonverbal behavior in employment interviews and video resumes.

2000–2006 **Masters** of Microengineering — EPFL, Switzerland
Spec. in robotics and autonomous systems
Master project: vision-based algorithm for the 3D simultaneous localization and mapping of a moving camera.
GPA: 5.82/6.

2003–2004 **Exchange year** Electrical Engineering — McGill University, Montréal, Canada
Exchange year for the two last semesters of my BSc.

experience

2011–Now **Idiap Research Institute - Social Computing Group** — Martigny, Switzerland
Research Assistant and PhD Student
During my PhD, I investigated the use of computational methods to automatically infer hirability in employment interviews and video resumes.

- Data Collections
  - SONVB: contributed to the setup of the sensor setting for the collection of 62 employment intervews.
  - Video resumes: collected 900+ video resumes from YouTube.
  - UBImpressed: led design, sensor setup, and collection of job interview and reception desk scenarios in hospitality (in progress).
- Data Processing
  - Used, developed and evaluated a variety of available methods to extract behavioral cues from audio, video, and speech transcripts.
  - Used and evaluated several supervised machine learning methods to infer social variables.
  - Designed and evaluated crowdsourcing experiments for the annotation of first impressions.
- Publications
  - Published one journal and two conference papers as first author.
  - Co-authored three journal papers and three conference papers spanning the literature in psychology and computing.

2007–2009 **EPFL - Ecological Engineering Lab** — Lausanne, Switzerland
Research assistant

- Led the development of a vision-based system for the monitoring of sewer networks.
- Co-authored a three-year CTI-funded project proposal.
- Published one journal paper as first author and co-authored one journal paper.

2000–2006 **Personal Tutor** — Nyon and Lausanne, Switzerland
Fields: mathematics and physics
Level: from high-school to first-year undergraduate.

awards

2014 **Idiap student paper award** Idiap Research Institute, Martigny, Switzerland
This award is given yearly to one Idiap PhD student for his outstanding paper.

interests

professional: multidisciplinary research, social computing, data analysis, multimedia processing, social media.
personal: playing music (guitar at home, bass guitar and vocals with a band), collecting LPs, travelling (9 months backpacking in 2009-2010), seasonal sports.

publications

as first author

Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior
L. S. Nguyen, D. Frauendorfer, M. Schmid Mast, D. Gatica-Perez
IEEE Transactions on Multimedia 16.4 (2014) pp. 1018–1031. 2014

Multimodal analysis of body communication cues in employment interviews
L. S. Nguyen, A. Marcos-Ramiro, M. Marrón Romera, D. Gatica-Perez
Proceedings of the ACM International Conference on Multimodal Interaction (ICMI), 2013

Using self-context for multimodal detection of head nods in face-to-face interactions
L. S. Nguyen, Jean-M. O. D. Gatica-Perez
Proceedings of the ACM International Conference on Multimodal Interaction (ICMI), 2012

as co-author

Nonverbal social sensing in action: Unobtrusive recording and extracting of nonverbal behavior in social interactions illustrated with a research example
D. Frauendorfer, M. Schmid Mast, L. S. Nguyen, D. Gatica-Perez
Journal of Nonverbal Behavior. Special Issue: Contemporary Perspectives in Nonverbal Research 38.2 (2014) pp. 231–245. 2014

Reliability and validity of nonverbal thin slices in social interactions
N. A. Murphy, J. A. Hall, M. Schmid Mast, M. A. Ruben, D. Frauendorfer, D. Blanch-Hartigan, D. L. Roter, L. S. Nguyen
Personality and Social Psychology Bulletin In press (2014). 2014

Social sensing for psychology: Automated interpersonal behavior assessment
M. Schmid Mast, D. Gatica-Perez, D. Frauendorfer, L. S. Nguyen, T. Choudhury
Current Directions in Psychological Science In press (2014). 2014

The Vernissage Corpus: A conversational human-robot-interaction dataset
D. B. Jayagopi, S. Sheikhi, D. Klotz, J. Wienke, J.-M. Odobez, S. Wrede, V. Khalidov, L. S. Nguyen, B. Wrede, D. Gatica-Perez
Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2013

Body communicative cue extraction for conversational analysis
A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, L. S. Nguyen, D. Gatica-Perez
Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2013

A semi-automated system for accurate gaze coding in natural dyadic interactions
K. A. Funes Mora, L. S. Nguyen, D. Gatica-Perez, J.-M. Odobez
Proceedings of the ACM International Conference on Multimodal Interaction (ICMI), 2013

pre-doctoral

Flow measurements in sewers based on image analysis: automatic flow velocity algorithm
D. Jeanbourquin, D. Sage, L. S. Nguyen, B. Schaeli, S. Kayal, D. A. Barry, L. Rossi
Journal of Water Science & Technology 64.5 (2011) pp. 1108–1114. 2011

Vision-based system for the control and measurement of wastewater flow rate in sewer systems.
L. S. Nguyen, B. Schaeli, D. Sage, S. Kayal, D. Jeanbourquin, D. A. Barry, L. Rossi
Journal of Water Science & Technology 60.9 (2009) pp. 2281–2289. 2009

HydroPix: a vision-based tool for the control of hydraulic structures in sewer systems
L. S. Nguyen, D. Sage, B. Schaeli, S. Kayal, L. Rossi, D. A. Barry
Geophysical Research Abstract, European Geoscience Union conference, 2009