I Would Hire You in a Minute: Thin Slices of Nonverbal Behavior in Job Interviews

Laurent Son Nguyen
Idiap Research Institute
École Polytechnique Fédérale de Lausanne
Switzerland
lnguyen@idiap.ch

Daniel Gatica-Perez
Idiap Research Institute
École Polytechnique Fédérale de Lausanne
Switzerland
gatica@idiap.ch

ABSTRACT

In everyday life, judgments people make about others are based on brief excerpts of interactions, known as thin slices. Inferences stemming from such minimal information can be quite accurate, and nonverbal behavior plays an important role in the impression formation. Because protagonists are strangers, employment interviews are a case where both nonverbal behavior and thin slices can be predictive of outcomes. In this work, we analyze the predictive validity of thin slices of real job interviews, where slices are defined by the sequence of questions in a structured interview format. We approach this problem from an audio-visual, dyadic, and nonverbal perspective, where sensing, cue extraction, and inference are automated. Our study shows that although nonverbal behavioral cues extracted from thin slices were not as predictive as when extracted from the full interaction, they were still predictive of hirability impressions with R^2 values up to 0.34, which was comparable to the predictive validity of human observers on thin slices. Applicant audio cues were found to yield the most accurate results.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: User/Machine Systems—*Human factors*

Keywords

Social computing; job interview; thin slices; nonverbal behavior; hirability; first impressions; multimodal interaction

1. INTRODUCTION

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 [5]. Short segments of interactions, typically under five minutes, are commonly referred to as thin slices [5]. Surprisingly, such minimal displays of behavior can be predictive of social constructs (*e.g.*, personality, competence) and outcomes (*e.g.* teacher ratings) [5]. As an extreme example of thin-slicing, inferences of competence by naïve raters based on simple photographs were strongly correlated with election outcomes [31].

Used in virtually every organization for the personnel selection process, job interviews are a prototypical situation in which first impressions play a crucial role; indeed, the hiring decision is most often based on how the job applicant was perceived by the recruiter. Employment interviews consist of an interpersonal interaction between at least two protagonists (the applicant and one or more interviewers). Because protagonists are strangers, recruiters have access to very little information (usually, the verbal and nonverbal elements of the interaction, as well as the resume), which makes job interviews close to what psychologists refer to as *zero-acquaintance interactions* [4], and nonverbal behavior is known to play a key role in the formation of first impressions [17].

In this work, we investigate the use of job interview thin slices for the task of automatically inferring hirability first impressions. To the best of our knowledge, no computational study has examined the role of thin slices in this type of interactions. We believe that job interviews are an interesting setting to investigate the interplay between thin slices and nonverbal behavior. This study 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?
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. Our study shows that although nonverbal behavioral cues extracted from thin slices were not as predictive as when extracted from the full interaction, they were still predictive of hirability impressions, with R^2 values up to 0.34. Automatically extracted nonverbal cues were found to yield comparable results to slice annotations in the task of predicting hirability impressions based on the full interview.

We believe that the answers to these questions could be applied to real-life application scenarios such as automated

ICMI'15, November 09 – 13, 2015, Seattle, WA, USA

ACM 978-1-4503-3912-4/15/11 ...$15.00.
http://dx.doi.org/10.1145/2818346.2820760.

interview training systems or HR assistive services by providing insights on (1) what questions and/or what moments in the interviews are most predictive of the outcome, (2) the amount of behavioral information necessary to obtain an accurate inference, and (3) what cues are predictive and how consistent they remain when thin-sliced.

This paper is structured as follows. In Section 2, we discuss the work related to thin slices and job interviews both in the psychology and computing domains. In Section 3, we present the dataset and the annotations used for this study. In Section 4, we discuss the nonverbal cues automatically extracted for both the applicant and the interviewer. In Section 5, we present the automatic inference task and discuss the results. We finally conclude in Section 6.

2. RELATED WORK

2.1 Related work in psychology

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 [3]. 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) [3]. 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 [5, 3].

Employment interviews have been a major research topic in social psychology for decades. In particular, previous works have investigated the reliability (*i.e.*, the level of agreement among judges for rating applicants) [15] and validity (*i.e.*, the amount of correlation between interview ratings and job performance) [27] of employment interviews, as well as the relationship between high-level social variables (*e.g.*, personality traits, general intelligence) and job performance [28, 7]. Particular attention has been put on the impact of the applicant's nonverbal behavior on the job interview outcome, and results have shown that applicants who employed more eye contact, smiled more, nodded more, produced more facial expressions, and were more oriented towards the interviewer were generally more likely to be hired, and were perceived as more competent, motivated, and more successful than applicants who did not [13, 25, 6, 16].

Surprisingly, only a few studies have investigated the effect of impressions from thin slices on job interview outcomes. In an unpublished Master's thesis [14], 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 [30] 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.

2.2 Related work in computing

Several computational studies have examined the use of thin slices in contexts similar to job interviews. The work in [11] studied the relationship between nonverbal behavior and outcomes in a simulated dyadic negotiation scenario, focusing on the first five minutes of the interaction. In [8], 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 [19], personality [9, 26], attraction [19], emergent leadership [29], or individual performance [18] in face-to-face interactions. In most of these works [19, 8, 18], 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 [29] 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.

Despite the importance of job interviews in the personnel selection process, to our knowledge only a few studies have analyzed this type of interactions from a computational perspective. In previous works, we have investigated the use of automatically extracted nonverbal cues (speaking turns, prosody, head nods, visual activity) to infer five types of hirability variables in a dataset of real job interviews [21]. In another study, we examined the relationship between visual body activity and personality and hirability using a mixture of automatically and manually extracted features [22]. Naim *et al.* [20] collected a dataset of 69 internship-seeking students participating to mock interviews, and extracted nonverbal and lexical features for the prediction of various social variables (*e.g.*, hiring recommendation, engagement, friendliness) and perceived behaviors (*e.g.*, smile, eye contact, speaking rate). Last, Chen *et al.* [10] proposed a multimodal method to recognize job applicants' affective states from 20 acted video interviews. To our knowledge, no previous study has investigated employment interviews from a thin-slice perspective.

3. DATASET

3.1 Job interview dataset

We used the dataset of 62 real job interviews collected in [21]. Applicants were applying for a paid marketing job in which they had to convince people on the street to participate to psychology studies (USD200 for four hours of effective work). The interviews were dyadic, where one applicant and one interviewer were seated at opposite sides of a table. All interviews were conducted by the same person, a researcher in organizational psychology.

No binary hiring outcome was given as such; all applicants were hired for the job regardless of their performance during the interview. However, applicants were performing at their best because they were believing that the interview outcome was depending on their performance. The interviewer did not rate the applicants, but video recordings of the interviews were sent to human resource professionals for rating (see Sections 3.3.1 and 3.3.2 for details).

The job interviews were structured, meaning that the sequence of questions/answers was constant across interviews, ensuring that comparisons could be made across job applicants. The interview structure was the following:

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

The interview dataset contains audio-video data for both the applicant and the interviewer. RGB video was recorded at 26.6 frames per second with a 1280×960 resolution. Cameras were nearly frontal, filming the upper part of the body. For audio, a Microcone microphone array was used, which provides semi-automatic speaker segmentation in addition to recording audio at 48kHz [1].

3.2 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 question *vs.* answer segments, we further annotated the timing of questions and answers. The annotations of timings were completed by one of the authors. In summary, three thin-slice cases were used: whole-slice, question-only, and answer-only (see Table 1).

Statistics of slice durations are shown in Table 2. We first observe that the longest slice (whole-slice) was 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).

3.3 Hirability impressions

3.3.1 Hirability impressions from the full interview

Hirability impressions were annotated by a pool of five human resources (HR) professionals, who watched the full videos of the applicant including the audio track. All interviews were annotated in total by three raters. The HR professionals were asked to rate the applicants' overall performances, and gave a score on a 10-point Likert scale, where 1 was the minimum score, and 10 the maximum score. To avoid bias, no specific instruction was given on what behaviors to focus on so that they could focus on their own internal representations.

Table 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

Inter-rater agreement proved to be high among HR professionals, with $ICC(1,1) = 0.501$ and $ICC(1,k) = 0.751$, using the intraclass correlation coefficient. Four other hirability scores related to questions 4-7 of the interview were annotated, but were not used in this study as they were highly correlated with the hiring decision score ($r \in [0.610, 0.916]$, using Pearson's correlation coefficient).

3.3.2 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, *i.e.* including question and answer), including both audio and video. 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. In order to avoid bias, the professional raters were different from the ones who annotated the full interviews (Section 3.3.1), but had similar backgrounds.

To assess agreement among 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 1 (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 reliability of judges in forming first 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 $r = 0.85$.

3.3.3 Observer predictive validity

We then computed the pairwise correlations between slice and full interview annotations, using Pearson's correlation coefficient. Slice-full correlations are displayed in Figure 1 (2). In social psychology related works [3], this measure is often used to quantify the observer predictive validity of the slice, which corresponds to the amount of information that an unacquainted judge can infer from a short excerpt of behavioral stream compared to the full interaction. 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

Table 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

(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 thin-slice 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 can be used for the comparison with automated methods in a regression task. The R^2 values for each slice are displayed in Figure 1 (3). We observe that the obtained R^2 values range from $R^2 = 0.06$ (communication skills) to $R^2 = 0.47$ (stress resistance).

4. BEHAVIORAL FEATURES

4.1 Extracted features

In order to obtain a behavioral representation of the interaction between the applicant and the interviewer during interview slices, we automatically extracted nonverbal cues from the audio and visual modalities for both protagonists. The extracted nonverbal cues were based on their relevance in the nonverbal communication literature [12, 16, 17] and the available computational tools. Other behavioral cues (*e.g.* gaze, facial expressions, verbal content) could also be extracted to get a more complete representation of the interaction, but we limited ourselves to the set of features listed below. The analysis of other behavioral cues will be the focus of future work.

All features were extracted for the eight interview questions (see Section 3) and the three thin slice cases (see Table 1), and were normalized with respect to slice duration. Additionally, the features were also extracted for the full interview. The list of all extracted nonverbal features is displayed in the **supplementary material**.

4.1.1 Audio features

Speaking activity features such as fluency or speaking time were shown to have an impact on hirability ratings [12]. All speaking activity cues were based on the speaker segmentations provided by the Microcone device [1]. The following cues were extracted (in brackets, we list the statistics used as features):

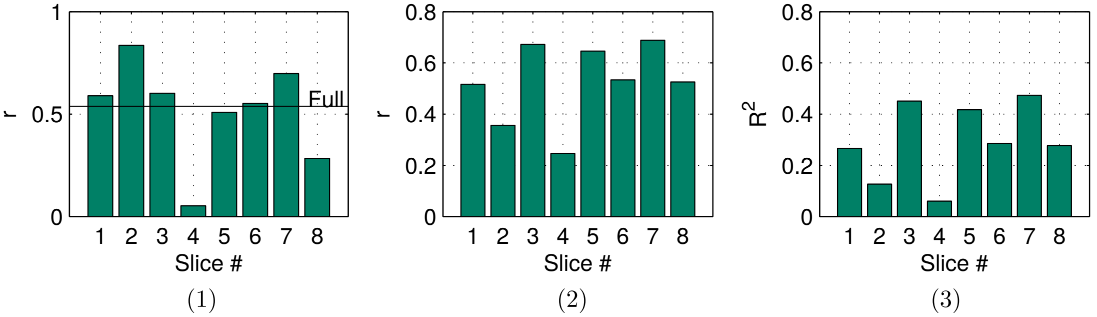
- *Speaking time* was obtained by summing all speaking turn durations in the slice of interest.
- *Speaking turns* were defined as speaking segments longer than 2 seconds in the slice of interest (number of turns, average, standard deviation, and maximum of turn durations).
- *Pauses* were defined as non-speaking gaps of duration smaller than 2 seconds (number of pauses).
- *Silence* was defined as moments when none of the protagonist was speaking (number of events and total duration).
- *Overlapping speech* was defined as events when both protagonists were speaking at the same time (number of events and total duration).
- *Short utterances* were defined as speaking segments of duration smaller than 2 seconds (number of events).
- *Audio backhannels* were defined as short utterances when the other person was speaking (number of events).

Prosody relates to the variations of tone that accompany speech and includes pitch (voice fundamental frequency), speaking rate (speed at which syllables are burst), and energy (voice loudness) [17]. These cues were shown to be correlated with hirability ratings in psychology studies [12]. We used the speech feature extraction code [2] from the MIT Media Lab. For speech energy, pitch, and voiced rate, we extracted the mean, standard deviation, minimum, maximum, entropy, median, and quartiles.

4.1.2 Visual features

Overall visual motion quantifies the amount of visual information displayed by the protagonist throughout the slice of interest and is an indicator of kinetic expressiveness. To compute it, we used Weighted Motion Energy Images (WMEI) developed in [9] which summarizes the motion during a time segment into a single grayscale image, each pixel intensity corresponding to the visual activity at this location. Statistics from WMEI images were then extracted: mean, median, standard deviation, minimum, maximum, entropy, and quartiles.

Head-region visual motion quantifies the amount of motion displayed in the face region. We used the parametric optical flow method developed in [24] located in the face bounding box to extract the head-region optical flow. Statistics from this time-series were then computed: mean, median, standard deviation, maximum, entropy, and quartiles.



(1) (2) (3)

Figure 1: (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).

Head nods were extracted using the method developed in [23]; total nodding time and number of events were then extracted from the obtained binary time-series.

4.1.3 Multimodal features

Visual backchannels were defined as an event where one protagonist was nodding while the other person was speaking; they are often used to signal agreement and enhancing communicative attention [17]. Total backchanneling time and number of events were used as features.

Nodding while speaking were defined as events when a protagonist was nodding while he was speaking. In terms of communicative behavior, nodding while speaking can *inter alia* be used to elicit feedback from the listener [17]. Total nodding-while-speaking time and number of events were computed.

4.2 Correlation analysis

Pairwise correlations between single behavioral cues extracted from thin slices and hirability impressions obtained from the full interview were computed. Due to space constraints, the correlation values are not displayed here but can be found in the **supplementary material**.

4.2.1 Applicant cues

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 [17].

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 ranged 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 ($r = 0.05$).

4.2.2 Interviewer cues

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 findings of [21]: 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.

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, due to the relative sparsity of head nods. This issue of temporal stability needs to be examined in more detail.

5. INFERENCE

5.1 Experiments

We defined the prediction task as a regression task, where the goal was to infer the hirability scores annotated from the full interview, using behavioral features extracted from thin slices as input. 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.

For inference, several regression techniques were tested (ridge regression, random forest, LASSO, ordinary least squares). Ridge regression with no dimensionality reduction was found to consistently yield the best prediction accuracies, therefore results obtained with other methods are not presented here. 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, highly skewed features ($skewness > 1$) were transformed using log-transformation ($z = log(1 + x)$, where x and z denote the original and log-transformed feature, respectively); also, all behavioral features were normalized using the z-score, such that they had zero mean and unity variance.

We ran the inference task for all eight interview slices and for all three thin slice case (see Table 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). 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, which accounts for the amount of total variance explained by the model under analysis; it is a metric often used in both psychology and social computing when dealing with regression tasks. Note that the use of adjusted R^2 often used in psychology is unnecessary here because we use a cross-validation framework which by construction separates the training from the test sets; indeed, the purpose of the adjusted R^2 is to account for an increase in the number of predictors when fitting a ordinary least-squares model to a variable.

5.2 Results and discussion

Q1: Prediction from thin slices

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 2. 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 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.

Q2: Most predictive slices

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.

Q3: Predictive cues

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 2), 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 4.2, 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.

Q4: Interview question vs. answers

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 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.

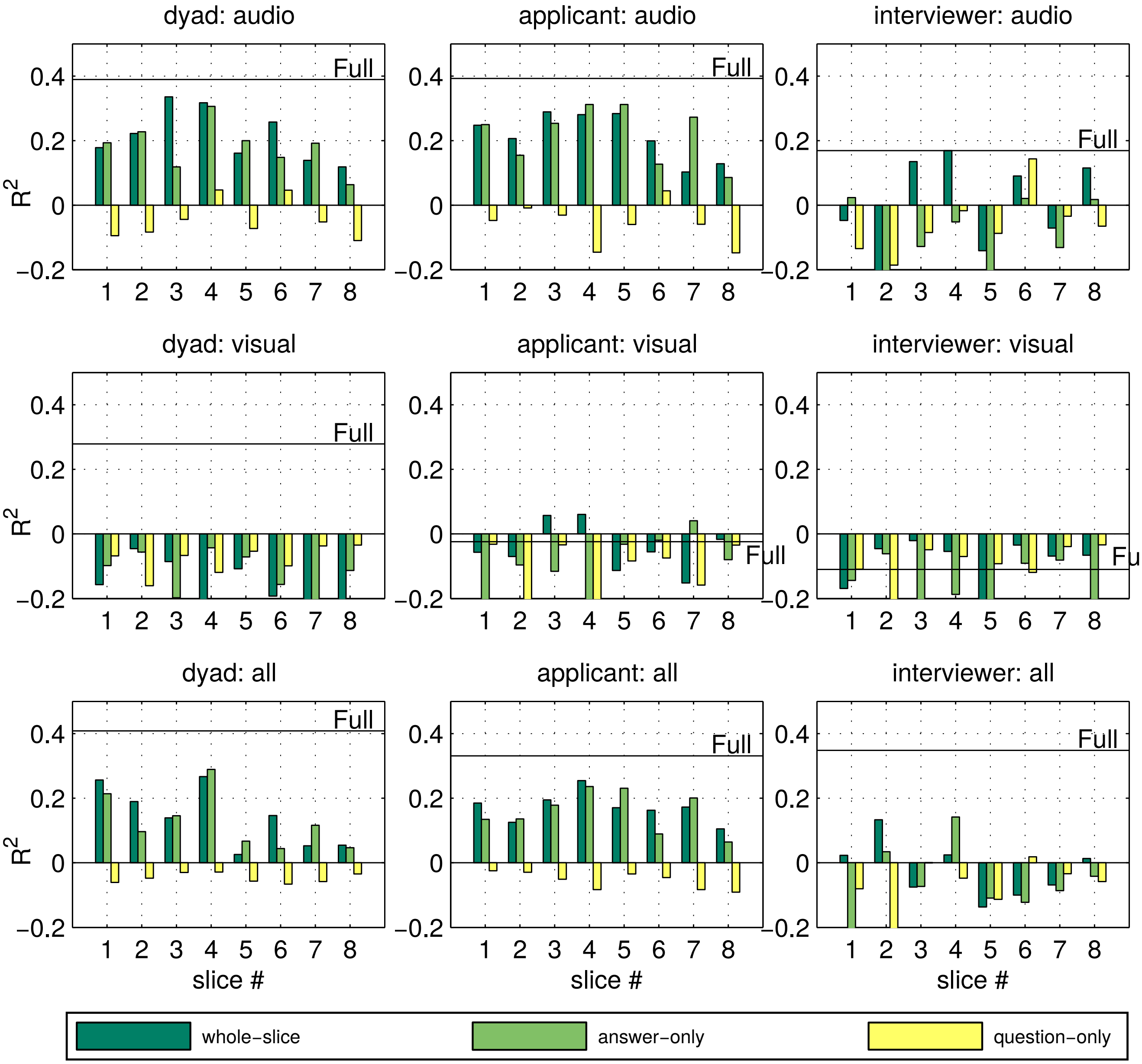


Figure 2: R^2 **results from thin slices for person-based feature groups (columns), modality-based feature groups (rows), and thin slice cases (dark green for whole-slice, light green for answer-only, and yellow for question-only). The solid line refers to the prediction result obtained using the full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

6. CONCLUSION

We analyzed thin slices of job interviews, where slices were defined by the specific questions of the interview structure. To our knowledge, this work constitutes the first computational attempt at understanding the effect of thin-slicing in job interviews. Among the main results, we found 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. Features extracted from thin slices were found to be not as predictive as the full interview, but 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. These results align with the previous findings in social psychology stating that nonverbal behavior plays an important role in the formation of first impressions, especially when the amount of information is low [17].

No slice clearly 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. We also examined the accuracy stemming from person- and modality-based feature groups, and applicant audio features yielded the most accurate results, while visual cues were quite unpredictive of hirability. One possible explanation for these results is the relatively limited number of behavioral channels analyzed in this work. Other types of behaviors such as gaze, gestures, facial expressions, smiles, and verbal content were shown to play an important role in the impression formation process [17] and might shed light on Q2, what questions/slices are the most predictive. This will be investigated in detail in future work.

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.

Beyond employment interviews, one of the challenges in thin slice research is the amount of temporal support necessary for each behavioral feature to be predictive of the outcome of the full interaction or the social construct of interest. In other words, some cues require to be aggregated over a longer period than others. Here, this was the case of the interviewer's nodding behavior which was predictive of hirability when extracted from the full interaction, but not from thin slices. To our knowledge, apart from the correlation between a thin-sliced feature and its full-interaction counterpart, no metric assessing the necessary amount of temporal support for a given feature exists; we believe that this would be beneficial for the field.

Acknowledgments

This work was funded by the UBIMPRESSED project of the Sinergia interdisciplinary program of the Swiss National Science Foundation (SNSF). Special thanks to Marianne Schmid-Mast and Denise Frauendorfer (Université de Lausanne) for the collaboration on experimental design and data collection.

7. REFERENCES

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

[2] Speech feature extraction code [online]. Available: http://groupmedia.media.mit.edu/data.php.

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

[4] N. Ambady, M. Hallahan, and R. Rosenthal. On Judging and Being Judged Accurately in Zero-Acquaintance Situations. *J. of Personality and Social Psychology*, 1995.

[5] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 1992.

[6] N. Anderson and V. Shackleton. Decision making in the graduate selection interview: A field study. *J. of Occupational Psychology*, 1990.

[7] M. Barrick and M. Mount. The Big Five personality dimensions and job performance: A meta-analysis. *J. of Personnel Psychology*, 1991.

[8] L. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe. Please, tell me about yourself: automatic personality assessment using short self-presentations. *Proc. ICMI*, 2011.

[9] J. Biel, O. Aran, and D. Gatica-Perez. You are known by how you vlog: Personality impressions and nonverbal behavior in YouTube. In *Proc. ICSWM*, 2011.

[10] L. Chen, M. Martin, and M. Ma. An initial analysis of structured video interviews by using multimodal emotion detection. In *Proc. Workshop on Emotion Representations and Modelling for HCI Systems*, 2014.

[11] J. Curhan and A. Pentland. Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes. *J. of Applied Psychology*, 2007.

[12] T. DeGroot and S. Motowildo. Why visual and vocal interview cues can affect interviewers' judgments and predict job performance. *J. of Applied Psychology*, 1999.

[13] R. Forbes and P. Jackson. Non-verbal behaviour and the outcome of selection interviews. *J. of Occupational Psychology*, 1980.

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

[15] A. Huffcut, J. Conway, P. Roth, and N. Stone. Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *J. of Applied Psychology*, 2001.

[16] A. Imada and M. Hakel. Influence of nonverbal communication and rater proximity on impressions and decisions in simulated employment interviews. *J. of Applied Psychology*, 1977.

[17] M. Knapp and J. Hall. *Nonverbal communication in human interaction.* Wadsworth, Cengage Learning, 2009.

[18] B. Lepri, N. Mana, A. Cappelletti, and F. Pianesi. Automatic prediction of individual performance from "thin slices" of social behavior. In *Proc. ICMI*, 2009.

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

[20] I. Naim, M. Tanveer, D. Gildea, and M. E. Hoque. Automated prediction and analysis of job interview performance: The role of what you aay and how you say it. In *Proc. FG*, 2015.

[21] L. Nguyen, D. Frauendorfer, M. Schmid Mast, and D. Gatica-Perez. Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Trans. on Multimedia*, 2014.

[22] L. Nguyen, A. Marcos-Ramiro, M. Marrón Romera, and D. Gatica-Perez. Multimodal analysis of body communication cues in employment interviews. In *Proc. ICMI*, 2013.

[23] L. Nguyen, J.-M. Odobez, and D. Gatica-Perez. Using self-context for multimodal detection of head nods in face-to-face interactions. In *Proc. ICMI*, 2012.

[24] J.-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *J. of Visual Communication and Image Representation*, 1995.

[25] C. Parsons and R. Liden. Interviewer perceptions of applicant qualifications: A multivariate field study of demographic characteristics and nonverbal cues. *J. of Applied Psychology*, 1984.

[26] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro. Multimodal recognition of personality traits in social interactions. In *Proc. ICMI*, 2008.

[27] R. Posthuma, F. Morgeson, and M. Campion. Beyond employment interview validity: A comprehensive narrative review of recent research and trends over time. *J. of Personnel Psychology*, 2002.

[28] S. Rothmann and E. Coetzer. The big five personality dimensions and job performance. *J. of Industrial Psychology*, 2003.

[29] D. Sanchez-Cortes, O. Aran, M. Schmid Mast, and D. Gatica-Perez. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Trans. on Multimedia*, 2012.

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

[31] A. Todorov, A. Mandisodza, A. Goren, and C. Hall. Inferences of competence from faces predict election outcomes. *Science*, 2005.

Appendix for the paper: "I would hire you in a minute: Thin slices of nonverbal behavior in job interviews"

Laurent Son Nguyen
Idiap Research Institute
École Polytechnique Fédérale de Lausanne
Switzerland
lnguyen@idiap.ch

Daniel Gatica-Perez
Idiap Research Institute
École Polytechnique Fédérale de Lausanne
Switzerland
gatica@idiap.ch

Appendix I: List of extracted nonverbal cues

In this appendix, we list all the extracted behavioral features used for the study. All features were extracted for all slices (and thin slice cases - question-only, answer-only, and whole-slice), as well as for the full interview. With the exception of silence/speech overlap, which refer to both protagonists jointly, all behavioral features were extracted for both the applicant and the interviewer. For details about the method used to extract these cues, please refer to the main paper. Table 1 and 2 display the list of audio features and visual features, respectively.

Table 1: List of extracted audio features.

Silence/speech overlap
Number of silent events
Time silent
Number of overlapping speech events
Time of overlapping speech
Speaking turns
Speaking time
Number of turns
Average turn duration
Maximum turn duration
Number of short utterances
Short utterances time
Number of pauses
Number of audio backchannel
Audio backchannel time
Prosody
Mean pitch
Pitch STD
Median pitch
Pitch lower quartile
Pitch upper quartile
Maximum pitch
Mean energy
Energy STD
Median energy
Energy lower quartile
Energy upper quartile
Maximum energy
Mean voiced rate
Voiced rate STD
Median voiced rate
Voiced rate lower quartile
Voiced rate upper quartile
Maximum voiced rate

Table 2: List of extracted visual features.

Head visual motion
Mean head horizontal visual motion
Horizontal head visual motion STD
Median horizontal head visual motion
Maximum horizontal head visual motion
Mean vertical head visual motion
Vertical head visual motion STD
Median vertical head visual motion
Maximum vertical head visual motion
Mean magnitude of head visual motion
Magnitude of head visual motion STD
Median magnitude of head visual motion
Maximum magnitude of head visual motion
Nodding
Number of nods
Nodding time
Number of visual backchannel events
Visual backchanneling time
Number of nodspeak events
Nodspeak time
Overall motion (WMEI)
WMEI horizontal center of mass
WMEI vertical center of mass
WMEI mean
WMEI median
WMEI standard deviation
WMEI lower quartile
WMEI upper quartile
WMEI maximum
WMEI ratio of non-zero pixels

Appendix II: Correlation

Tables 3 and 4 display the list of applicant and interviewer cues significantly correlated with the hirability impression obtained from annotations of the full interview. Note that only cues showing significant correlations ($p < .05$) are displayed. Please refer to the main paper for the discussion about these results.

Table 3: Applicant nonverbal cues extracted from thin slices (whole-slices) and from the full interview significantly correlated with the hirability impressions annotated from the full interview ($p < .05$, $\dagger p < .005$). $N = 62$.

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 4: Interviewer nonverbal cues extracted from thin slices (whole-slices) and from the full interview significantly correlated with the hirability impressions annotated from the full interview ($p < .05$, $\dagger p < .005$). $N = 62$.

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			