Automated Video Interview Judgment on a Large-Sized Corpus Collected Online

Lei Chen*, Ru Zhao†, Chee Wee Leong*, Blair Lehman*, Gary Feng*, Mohammed (Ehsan) Hoque†
*Educational Testing Service (ETS)
Princeton, NJ
{LChen, CLeong, BLehman, GFeng}@ets.org
†Computer Science
University of Rochester
Rochester, NY
{rzhao2, mehoque}@cs.rochester.edu

***Abstract*—Online video-based job interviews are becoming very popular in the screening of potential employees. In this study, we collected a corpus of 1891 monologue job interview videos (63 hours in duration) from 260 online workers. These videos were annotated for personality traits and hiring recommendation score by experts from a major assessment company. We proposed a unified method of automatic analysis that consists of using clustering to convert continuous audio/video analysis output to discrete pseudoword documents, and then applying modern text classification methods to process speech content, prosody and facial expressions. Our experiments showed that using what the interviewees say (i.e., spoken text), we can predict their personality traits such as openness, conscientiousness, extraversion, agreeableness, and emotional stability with an F-measure of 0.8 or better, while we get an F-measure of 0.6 in predicting hiring recommendation score. Prosody and facial expressions added limited usefulness on interview judgments and need further investigation.**

1. Introduction

In recent years, online video-based interviews have been increasingly used in the hiring processes [1] , and brought many benefits to both interviewers and interviewees, including the convenience of off-line reviewing and decision making by human resources (HR) staff, which in turn enables HR staff to assess multiple job applicants in a short time window. It also opens the door for automated performance analyses to assist in initial HR decision making and possibly also reduce human biases.

During interviews, interviewees must effectively broadcast their enthusiasm and expertise through their multimodal behaviors, such as speech content, prosody, gaze direction, facial expressions, and other nonverbal cues, in a limited amount of time [2], [3], [4]. The success or failure of the interviewee's effort is traditionally assessed subjectively by the interviewer, either through a holistic impression or quantitative ratings. The validity and reliability of these assessments are subject to much debate [5]. An emerging alternative to the traditional human-only interview assessment model is to augment human judgment with automated assessment of interview performance based on social signal processing (SSP) [6], [7].

Compared to the previous research that will be briefly surveyed in Section 2, our contributions in this paper can be summarized as follows. First, we collected a corpus of 1891 monologue videos (63 hours in overall duration) via a crowd-sourcing approach. Second, we proposed a unified automatic rating approach to effectively process multimodal behaviors.

The remainder of the paper is organized as follows. Section 2 reviews the previous research on interview judgment using SSP. Section 3 describes our large-sized monologue video interview corpus that was collected via a crowd-sourcing approach. Section 4 describes the machine prediction pipeline, including multimodal data processing steps and machine prediction. Section 5 reports on our experimental results. Lastly, Section 6 discusses our findings and plans for the next steps of our research.

2. Previous Research

The present study focuses on the automated scoring of interview video responses in a task in which the interviewee responds to a fixed set of standardized questions, also known as a structured interview (SI). Research from Industrial Organizational (I-O) psychology shows that SIs tend to produce more valid results than unstructured interviews [8]. A growing body of research examined interviewees' video responses to SI questions. For example, [9], [10] investigated the validity of using a webcam to collect job interview responses and reported "webcam test score manifested a significant and positive correlation with job placement success ($r = 0.26, p < .05$)".

Research efforts in developing automated video interview judgment systems have emerged [11], [12]. In [11], a multimodal corpus consisting of 62 interviews of candidates applying to a real temporary job was built. Each interview lasted approximately 11 minutes. Four interview questions measuring job-related skills in *communication*, *persuasion*, *conscientiousness*, and *coping with stress*, were

used. A Master's student majoring in I-O psychology rated each question on a 5-point Likert scale and also the entire interview on a 10-point scale for hiring recommendation. In addition, other questionnaires (e.g., related to the interviewee's personality) were also administered. From both the interviewees and the interviewers, various audio features (e.g., speaking activity, pauses, prosody, etc.) and visual behavior cues (e.g., head nods, smiling, etc.) were automatically extracted. Gazing and physical appearance were annotated manually. The authors also developed several cross-modality features using information from both audio and video channels jointly. Afterwards, these multimodal cues were used to predict five types of human-rated scores using different machine learning approaches.

Another study [12] focused on 138 audio-video recordings of mock interviews from internship-seeking students at Massachusetts Institute of Technology. The total duration of the recorded interviews is about 10.5 hours. Career counselors asked interview questions that were recommended by MIT Career Services to measure student applicants' behavioral and social skills. Sixteen 7-point Likert scale questions were used to rate the interviewees' performance, consisting of two questions on overall performance (*overall rating* and *recommended hiring*) and 14 questions targeting behavioral dimensions (e.g., *presence of engagement*). The ratings were conducted by counselors and Amazon Mechanical Turk (AMT) workers. The automatic analysis used the following multimodal cues: facial expressions, language (e.g., word counts, topic modeling), and prosodic information. The ground truth ratings were obtained by a weighted average over the ratings from 9 Mechanical Turker raters. These multimodal cues were fed into machine learning models to predict human ratings. For some interview traits (e.g., excitement, friendliness, and engagement), the automatic prediction showed correlation coefficients of 0.75 or higher with human ratings.

Research interests in the automatic judgment of video interviews have been increasing, and two new data sets appeared in 2016. In [13], both video interviews and face-to-face interviews were collected from 106 college students attending a university in India. Based on human ratings done by three external reviewers, the authors found that there was only a slight difference between the video and face-to-face interviews, providing support that video-based interviews can be used to replace face-to-face interviews in some controlled scenarios. In [14], [15], a total of 36 subjects participated in the study and each participant was required to answer 12 SI questions. Note that except [11], all these video interview data collections [12], [13], [14] did not have specific jobs to fill and measured generic job related skills.

The ROC Speak[1] system presents a framework from which we drew insight regarding crowd-sourcing a large corpus of video-based interviews. The ROC Speak system enables users to record video from their web browsers and immediately analyze the audiovisual signals to provide feedback on many nonverbal cues, such as pitch, volume, and movement [16]. Such a system can collect a large volume of data from users practicing public speaking in the comfort of their practice environment. For example, in [17], the ROC Speak system was deployed online for collecting 196 videos from 49 users practicing speeches in front of their computers. [17] was made possible by the ubiquitousness of the collection method. This provides evidence that we can use a similar collection method to enable the automatic assessment of video interviews.

Our brief survey suggests several limitations related to the prior research. Firstly, the corpora used in previous studies are limited in terms of their sizes, subject pools, and diversity. Most of these data sets contain fewer than 100 subjects, and these subjects are located in one place. Therefore, it will be hard to directly generalize the models developed on these data sets to real data consisting of a large number of subjects located in different places. Regarding multimodal sensing technology, the existing feature extraction methods lack of a unified framework. This paper is meant to address these two critical limitations.

3. Crowd-Sourcing Interview Collection

Based on many years of I-O psychology research (e.g., [18]), we decided to use structured interviews (SIs) and past-focused behavioral questions, in which the applicant is asked about how he or she has handled work-related situations in the past, in our data collection. Eight past-focused behavioral SI questions were used to assess 4 types of social skills valued in the workplace [19]: (a) *communication skills*, (b) *interpersonal skills*, (c) *leadership*, and (d) *persuasion and negotiating*. Figure 1 shows a question related to the leadership skills being asked about in our data collection.

Question: Please tell us about a work situation in which you were not the formal leader but tried to assume a leadership role. Please provide:

- details about the background of the situation,
- the behaviors you carried out in response to that situation,
- and what the outcome was.

Figure 1. A sample interview question measuring leadership skills

We used a technology similar to ROC Speak to collect behavioral data from participants remotely. A JavaScript-based video recording library—RecordRTC [20]—was used to record participants' video responses. Using a Chrome browser on a computer with a webcam and Internet connection, anyone can participate in the data collection by visiting our web-based collection interface[2].

At the beginning of the study, users were given a view of their webcam feed to adjust their webcam position and were asked to go to a quiet area before proceeding. Users were then given an interview question on the display

1. http://rocspeak.com

2. https://www.machinteraction.com/ETSstudy/

and one minute to prepare their answer. Following the one minute preparation time, users were notified by the display to answer the interview question as the computer started recording their webcam feed. The participants were given two minutes to respond to each interview question. A countdown timer was shown on the web page to help participants manage their allowed time. A WebM file was generated and stored inside the browser temporarily. Users then proceeded to prepare for the next interview question while the computer uploaded the stored video to the server in the background. Video was recorded in 480p resolution with 30 fps and audio was recorded in a mono channel with a $48K$ Hz sampling rate. This process was repeated until users answered all 8 interview questions. After each session, 8 video responses answering the SI interview questions were collected on the server. Separating video recording and uploading brings several advantages. Since the interviewees' browsers have more time to work on uploading videos, obtaining high quality videos becomes possible. Also, such a setup can enable participants whose Internet connection is slow to participate in data collection.

TABLE 1. COMPARISON OF OUR CORPUS WITH CORPORA IN THE PREVIOUS STUDIES

Work	Subj.	Dur. (min)	Type	Location
[11]	62	682	dialog	local (Switzerland)
[12]	138	630	dialog	local (USA)
[13]	106	1696	mono/dia-log	local (India)
[14]	36	753	monologue	local (USA)
this study	260	3784	monologue	on-line (USA)

Using the web-based video collection interface, we collected interview video responses from Mechanical Turk workers (Turkers) located in United States using the Amazon Mechanical Turk (AMT) marketplace. On our HIT page, we provided a description of the task and the consent form. After a participant accepted the consent form, he or she was transferred to the data collection web page. After recording, the Turker was compensated with 1 USD automatically and then an additional 9 USD if his or her videos met our standard.

We ultimately obtained valid videos from a total of 260 Turkers. Since Turkers are from different locations (any of the 50 states) and have a variety of profiles (gender, ethnicity, age), compared to the data sets in previous studies [11], [12], [13], [14], our subjects are more diverse. Also, different recording conditions (e.g., devices, lighting, and internet connections) appear in our data set, which provides working scenarios closer to future real applications. Table 1 compares our corpus with the corpora used in previous studies on the number of subjects (Subj.), total duration in minutes (Dur.), types of interviews, and locations for doing data collection.

We developed several rating scales based on raters' impressions of the video interview performance. In particular, we focused on personality traits and overall hiring recommendation as suggested in [19]. In this rating scenario, a set of statements was presented to the raters and the degree to which they agreed was measured on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree). Such a rating process, which requires no training or behavior anchors, is similar to those used in [12], [21]. The statements used for measuring personality traits contained adjectives (e.g., assertive, irresponsible, cooperative) that corresponded to the Big Five personality traits (*extroversion*, *agreeableness*, *conscientiousness*, *emotional stability*, and *openness to experience*), similar to the multiple-item measure used in [21]. Note that the personality factor-specific adjectives were selected for each type of interview question, resulting in 4 separate rating forms. Finally, raters were also asked to make a holistic judgment about hiring the participant for an entry-level office position. Five raters who have experience scoring writing essays and rating video performance at ETS rated video responses individually.

Since our aim is to develop automatic interview judgment systems, when analyzing human rating quality, we focused on video responses that could be successfully processed by the multimodal signal processing procedures that will be introduced in Section 4.1. Table 2 shows rater agreement on the $1,891$ videos (with a total duration of 63 hours) that will be used in our modeling experiment, including (a) Intra-class correlation coefficient (ICC) [22] (using the two-way random average measure of consistence, also known as $ICC(2,k)$), and (b) the minimal, maximal, and average correlation coefficient of individual raters' scores to the averaged scores. All scores' ICC values are higher than 0.75. This provides strong support for consistent and reliable ground truth for our large-sized video interview corpus. In our experiment, for each video, the averaged human scores from 5 raters were used as ground truth scores, following [22].

TABLE 2. ANALYSIS OF HUMAN RATING QUALITY: ICC, MAXIMUM (MAX), MINIMUM (MIN), AND MEAN OF INDIVIDUAL RATER'S SCORES' R TO THE AVERAGED ONE ON THE VIDEOS (N=1891) USED IN THE EXPERIMENT.

Category	ICC	R_{Min}	R_{Max}	R_{Mean}
Hiring recommendation	0.79	0.70	0.79	0.74
Agreeableness	0.91	0.73	0.94	0.86
Conscientiousness	0.95	0.79	0.96	0.92
Emotion stability	0.90	0.43	0.94	0.83
Extraversion	0.92	0.79	0.93	0.87
Openness	0.90	0.65	0.92	0.84

4. Methods

4.1. Multimodal Feature Extraction

When converting interviewee's audio responses to texts, we utilized the automatic speech recognition (ASR) service provided by IBM Bluemix platform. The ASR system performed well on our video interview responses. We checked several ASR outputs and noticed that the recognition accuracy was high enough for using the recognition content in follow-up studies.

We utilized the pyAudioAnalysis [23] package to extract the short-time acoustic features listed in Table 3[3]. In our analysis, each audio frame is 200 msec long.

We used OpenFace [24] for the following visual measurements: (a) tracking head poses, (b) tracking gaze directions, and (c) estimation of occurrence and intensity of Action Units. For each video frame with a width of 33.3 msec, head pose was extracted by using [25] and represented in two vectors, location and rotation, using world-coordination. Eye gaze directions were extracted by using [26]. Finally, occurrence conditions and intensity values of 18 AUs were generated by using [27]. Because they are a medium level representation that contributes different types of higher level behaviors (e.g., affects, engagement, etc.) we expect that AUs' temporal and spatial changes will carry useful visual information for rating interviews.

4.2. Models

After applying the diverse multimodal processing procedures listed in Section 4.1, for each interview video response, we obtained its recognized word string and audio/video analysis result vectors. Following [15], we converted the task of modeling audio/video analysis results into a 'text' classification task. In particular, we first obtained the mean of the audio/video analysis vectors along a time interval. In this study, the interval was 1.0 second for audio and 0.5 seconds for video in order to catch faster changes. Then, an unsupervised clustering method was applied on these averaged vector sequences to find K number of clusters. The entire sequence was then converted to a text string consisting of these discrete cluster numbers (called pseudo words hereafter). After converting audio and video analysis results to 'text documents', they can be modeled similar to ASR recognition outputs by using text feature extraction methods, for example, Bag of Words (BoW), n-gram, Doc2Vec [28], and so on. In this study, we used the Bag of Words (BoW) model to extract features. For a document (corresponding to one video response), the term frequency-inverse document frequency (TF-idf) of all word tokens were used as feature inputs. The entire process is summarized in Figure 2.

5. Experiment

On the 5 types of personality perception ratings and the holistic rating of hiring recommendation , we used the median value of scores in the entire dataset as a threshold to separate all instances into two labels, HIGH vs. LOW. Then, machine learning methods were used to build binary classifiers to distinguish interview videos into these two labels. Such a binary classification has many potential applications, such as identifying video responses that have a low score in order to provide constructive feedback when coaching potential job applicants, and pre-screening low-performance interview videos.

The entire data set (containing 1891 interview video responses) was split into Train ($n = 1591$) and Test ($n = 372$) sets. Note that when doing data splitting, we made sure that all the available video responses for any particular interviewee were in the same partition.A cross validation ($n = 5$) was run on the Train set to determine the best hyper parameters for the different steps, including clustering, text feature extraction, and classification. Then, the final model was trained using these hyper parameters on the Train set, and evaluated on the Test set. Regarding the measurement of binary classification performance, we used the macro F-1 measurement averaged on the two labels. A higher F-1 indicates more accurate binary classification.

The data modeling code was implemented using the scikit-learn Python package [29]. For clustering and BoW text feature extraction, we used the classes provided by the package directly. Regarding machine learning classifiers, we used two types of classifiers, including a Random Forest (RF) and a Support Vector Machine with a linear kernel (SVM). The modeling approach depicted in Figure 2 (highlighted in the box with dashed lines) shows that entire process contains three separate modules (i.e., clustering, feature extraction, and classification). Clearly, in order to obtain the best performance, an optimal combination of hyper parameters in these three steps needs be found. For example, K in the clustering step, the hyperparameters in a RF model (e.g., the number of trees in a forest n), the hyperparameters in a SVM model (e.g., penalty C), need to be determined jointly. We relied on the *pipeline* mechanism [30] provided in the scikit-learn package to conveniently determine the combination of optimal parameters.

Table 4 reports on the experimental results. Note that we only keep results better than chance performance. For personality perception classification tasks, text cues play a dominant role. Using spoken words provided by a modern cloud-based ASR system and their simple BoW feature formation, we can quite accurately classify high and low personality perceptions with an F-1 measure of about 0.8. Video cues were found to be useful for openness, conscientiousness, and agreeable personality perceptions, but with a low F-1 measure (of about 0.55). Regarding modeling hiring recommendation ratings on interviews, both text and audio cues show their usefulness with considerable F-1 measurements. (The F-1 for text is 0.66 while the F-1 for audio is 0.63). To investigate the power of their combination, we experimented with fusing these two information resources by both an early fusion (i.e., combining on the feature level) and a late fusion (i.e., combining on the prediction level). However, no further improvement was achieved after combining these two information resources in our experiment.

6. Discussion

In this paper, we present the automated analysis and prediction of job interview interview performance of 1891

3. The table was originally shown in https://github.com/tyiannak/pyAudioAnalysis/wiki/3.-Feature-Extraction

TABLE 3. ACOUSTIC FEATURES ON SHORT-TERM WINDOWS

ID	Feature Name	Description
1	Zero Crossing Rate	The rate of sign-changes of the signal during the duration of a particular frame.
2	Energy	The sum of squares of the signal values, normalized by the respective frame length.
3	Entropy of Energy	The entropy of the normalized energies of the sub-frames. It can be interpreted as a measure of abrupt changes.
4	Spectral Centroid	The center of gravity of the spectrum.
5	Spectral Spread	The second central moment of the spectrum.
6	Spectral Entropy	Entropy of the normalized spectral energies for a set of sub-frames.
7	Spectral Flux	The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8	Spectral Rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9-21	MFCCs	Mel Frequency Cepstral Coefficients

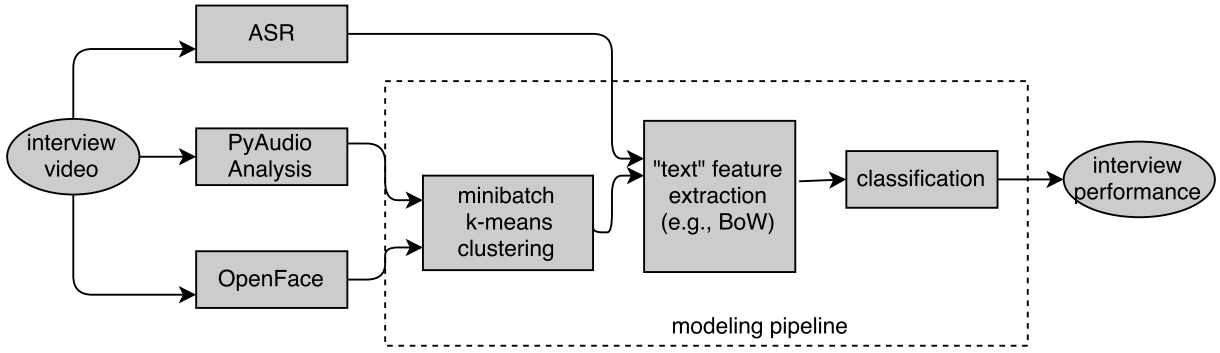

Figure 2. **Data flow diagram showing the process of using k-means clustering to obtain pesudowords from both audio and visual analysis results and applying a text classification method to evaluate job interview videos**

TABLE 4. EXPERIMENTAL RESULTS FOR DISTINGUISHING HIGH VS. LOW PERFORMANCE ON OVERALL HOLISTIC SCORE AND THE PERCEPTION OF 5 TYPES OF PERSONALITY TRAITS USING VARIOUS MULTIMODAL CUES. FOR EACH TYPE OF HUMAN SCORE, THE CLUSTER NUMBER (K) AND MACHINE LEARNING METHODS (SUPPORT VECTOR MACHINE (SVM) VS. RANDOM FOREST (RF)), CLASSIFICATION PRECISION, RECALL, AND F-1 MEASUREMENT, ARE REPORTED.

Modality	K	model	Precision	Recall	F-1
Openness					
Text		SVM ($C = 1.0$)	0.81	0.81	0.81
Video	50	RF ($n = 100$)	0.54	0.55	0.54
Conscientiousness					
Text		SVM ($C = 1.0$)	0.86	0.86	0.86
Video	20	RF ($n = 100$)	0.56	0.56	0.56
Extraversion					
Text		SVM ($C = 1.0$)	0.78	0.78	0.78
Agreeableness					
Text		SVM ($C = 1.0$)	0.84	0.84	0.84
Video	20	RF ($n = 50$)	0.56	0.56	0.55
Emotional Stablity					
Text		SVM ($C = 1.0$)	0.83	0.83	0.83
Hiring recommendation					
Text		SVM ($C = 1.0$)	0.67	0.66	0.66
Audio	20	SVM ($C = 1.0$)	0.64	0.63	0.63
T + A		early-fusion	0.65	0.65	0.65
T + A		late-fusion	0.63	0.63	0.63

videos that we collected through 260 Amazon Mechanical Turkers across United States. To enable such a large-scale data collection, an on-line video collection system was developed based on WebRTC technology to support convenient recording inside a Chrome browser without any software installation. The videos collected from Turkers were recorded using their laptops or smart phones in their own environments, which simulated a real online interview rating application. The insights reported in this paper are derived from a massive data set, which was carefully annotated by trained experts from a major assessment company. Our institutional review board (IRB) approval is such that we may share the videos and annotations to the affective computing and multimodal sensing community to further extend this research. In addition to collecting the data set, we proposed a general solution to analyze behaviors on the basis of speech content, prosody, and facial expressions. Using spoken words provided by ASR and the simple yet effective BoW feature representation, text-based classifiers of personality scores show high F-1 measurements. Audio and video models based on the proposed solution demonstrate their contributions (albeit small) to the classification tasks for different types of scores.

Given the several advantages of our corpus (i.e., diverse participants from multiple places, high quality videos, recorded in real life situations), we believe that further explorations are necessary, such as analyzing emotions shown in videos. Our study based on a large-sized video interview corpus suggests that the automatic video interview scoring based on the SSP technology is promising. With more investigations in future, such technique has a potential to play an active role in supporting HR decisions. In the near future, we plan to explore using a regression method to directly predict human rated scores, to investigate more sophisticated textual features, e.g., Doc2Vec [28], and to apply neural network based models. Also, our data-driven methods are influenced much by human ratings, including their decision biases. It is critical to maintain a high standard on assessment

fairness for both human and machine scorings. For example, a series of meta-data about interviewees (e.g., gender, age, ethnic group) need to be added in the corpus to evaluate the interview assessments' fairness.

Acknowledgments

The authors would like to thank Zydrune Mladineo for her great help running AMT data collection, Michelle Martin-Raugh and Harrison Kell for providing interview questions based on their I-O psychology expertise, Katrina Roohr and Kri Burkander for supporting the human rating work on this large amount of videos, Ubale Rutujia for providing programming support, and finally the human raters for providing high-quality ratings.

References

[1] A. Hiemstra and E. Derous, "Video resumes portrayed: Findings and challenges," *Employee recruitment, selection, and assessment: Contemporary issues for theory and practice*, pp. 44–60, 2015.

[2] R. J. Forbes and P. R. Jackson, "Non-verbal behaviour and the outcome of selection interviews," *Journal of Occupational Psychology*, vol. 53, no. 1, pp. 65–72, 1980.

[3] 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*, vol. 62, no. 3, pp. 295–300, 1977.

[4] 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*, vol. 69, no. 4, pp. 557–568, 1984.

[5] I. Nikolaou and J. K. Oostrom, *Employee Recruitment, Selection, and Assessment: Contemporary Issues for Theory and Practice*. Psychology Press, 2015.

[6] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 69–87, 2012.

[7] D. Gatica-Perez, "Signal processing in the workplace [social sciences]," *Signal Processing Magazine, IEEE*, vol. 32, no. 1, pp. 121–125, Jan 2015.

[8] T. DeGroot and J. Gooty, "Can nonverbal cues be used to make meaningful personality attributions in employment interviews?" *Journal of Business and Psychology*, vol. 24, no. 2, pp. 179–192, 2009.

[9] J. K. Oostrom, M. P. Born, A. W. Serlie, and H. T. van der Molen, "Webcam testing: Validation of an innovative open-ended multimedia test," *European Journal of Work and Organizational Psychology*, vol. 19, no. 5, pp. 532–550, 2010.

[10] J. Oostrom, M. Born, A. Serlie, van der M., and T. Henk, "A multimedia situational test with a constructed-response format: Its relationship with personality, cognitive ability, job experience, and academic performance." *Journal of Personnel Psychology*, vol. 10, no. 2, pp. 78–88, 2011.

[11] L. Nguyen, D. Frauendorfer, M. Mast, and D. Gatica-Perez, "Hire me: Computational Inference of Hirability in Employment Interviews Based on Nonverbal Behavior," *Multimedia, IEEE Transactions on*, vol. 16, no. 4, pp. 1018–1031, June 2014.

[12] I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque, "Automated prediction and analysis of job interview performance: The role of what you say and how you say it," in *Proc. of Automatic Face and Gesture Recognition (FG)*, 2015.

[13] S. Rasipuram, P. R. S. B., and D. B. Jayagopi, "Asynchronous video interviews vs. face-to-face interviews for communication skill measurement: a systematic study," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, Tokyo, Japan, 2016, pp. 370–377.

[14] L. Chen, G. Feng, M. Martin, C. W. Leong, K. Chris, S.-Y. Yoon, B. Lehman, H. Kell, and C. M. Lee, "Rating Monologue Video Interviews Automatically Using Multimodal Cues," in *Proceedings of the ISCA InterSpeech Conference*, San Francisco, CA, Sept. 2016.

[15] L. Chen, G. Feng, C. Leong, B. Lehman, M. Martin-Raugh, H. Kell, S. Yoon, and C. Lee, "Automated scoring of interview videos using Doc2Vec multimodal feature extraction paradigm," in *Proc. of ACM ICMI*, Tokyo, Japan, Nov. 2016.

[16] M. Fung, Y. Jin, R. Zhao, and M. Hoque, "ROC speak: semi-automated personalized feedback on nonverbal behavior from recorded videos," in *Proc. of UbiComp*, 2015.

[17] "ROC comment: automated descriptive and subjective captioning of behavioral videos," in *Proc. of UbiComp*, 2016.

[18] P. M. Wright, P. A. Lichtenfels, and E. D. Pursell, "The structured interview: Additional studies and a meta-analysis," *Journal of occupational psychology*, vol. 62, no. 3, pp. 191–199, 1989.

[19] A. I. Huffcutt, 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*, vol. 86, no. 5, pp. 897–913, 2001.

[20] A. B. Johnston and D. C. Burnett, *WebRTC: APIs and RTCWEB protocols of the HTML5 real-time web*. Digital Codex LLC, 2012.

[21] M. R. Barrick, G. K. Patton, and S. N. Haugland, "Accuracy of interviewer judgments of job applicant personality traits," *Personnel Psychology*, vol. 53, no. 4, pp. 925–951, 2000.

[22] G. G. Koch, "Intraclass correlation coefficient," *Encyclopedia of statistical sciences*, 1982.

[23] T. Giannakopoulos, G. Wakefield, E. R Hruschka, C. Fredouille, G. Friedland, and O. Vinyals, "pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis," *PLOS ONE*, vol. 10, no. 12, p. e0144610, dec 2015.

[24] T. Baltru, P. Robinson, L.-P. Morency *et al.*, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[25] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "Constrained Local Neural Fields for Robust Facial Landmark Detection in the Wild," in *2013 IEEE International Conference on Computer Vision Workshops*. IEEE, dec 2013, pp. 354–361.

[26] E. Wood, T. Baltrusaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling, "Rendering of Eyes for Eye-Shape Registration and Gaze Estimation," in *Proc. of ICCV*, 2015, pp. 3756–3764.

[27] T. Baltrusaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic Action Unit detection," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, may 2015, pp. 1–6.

[28] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," in *Proceedings of The 31st International Conference on Machine Learning*, 2014, pp. 1188–1196.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[30] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. C. Müller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. Vanderplas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," 2013.