Analytical feature extraction for Individual assessment through Machine Learning in Interface based Interview

Jasneet Singh Sawhney [1], Jasmeet Singh Sodhi [2] , Maninder Bir Singh Gulshan[3]
[1]Student, [2]Student, [3]Student
[1]Department of Electronics and Communication Engineering
[1]Dr. Akhilesh Das Gupta Institute of Technology and Management, New Delhi, India

Abstract : This paper presents a review on various techniques on predicting the interview performance autonomously using machine learning techniques and analyze the audio-visual features using an interface-based video interview by a virtual agent. Different scenarios are studied and compared for a better understanding of human behavior in case of the video interview. In this paper, the existing optimization techniques are reviewed according to their methodologies and advantages. Based on the review, research gaps are observed and the references are provided for further research work.

Index Terms **- Interface based video interviews, Automatic job prediction, Machine learning for job prediction, Nonverbal behavior prediction, Regression, Deep learning for audio-visual predictions.**

I. INTRODUCTION

Interviews are the most crucial and indispensable step a candidate goes through during the hiring process. Be it any position, from an intern to the CEO of the company, has to pass interview to prove his/her credibility for a particular position. An Interview process has both verbal as well as non-verbal features to analyze, and the result of the analysis can be the binary prediction like selected or not, or a better screening pipeline can be devised by doing a score based analysis.

Interviews can be very nerve-wracking, you want to make a great first impression and be able to tick all of the company's boxes, so that they either invite you for a second interview or offer you the job.

According to the Twin Employment and Training, results from a survey of 2,000 hiring managers found that 33% knew whether they would hire someone in the first 90 seconds. So, the first impression plays a crucial role in hiring [1].

Factors like smile, prosody, posture, head positions and hand movements also play a vital role in the interview process. Making proper eye contact, face movements, speaking rate (words per minute), loudness and clarity of voice are some added features, which if correctly mastered can boost the performance multifold.

Companies hire interns, job professionals, freelancers, and many other micro and macro scale workers, all through a basic interview process, which increases the value of this process by a great extent. Large companies offer bulk hiring which means screening hundreds of candidates in one go. At the time of writing this paper, there are 1634 companies listed on Naukri.com which are open for bulk hiring both onsite and remote. In bulk hiring, the main chunk of the cost of the interview includes the appointment of recruiters who can take interviews and screen the students. The capabilities and mindset of every recruiter are different and hence there is no standard which could be set even if the recruiters are given pre-decided parameters to judge on, then also there would be human error. This problem could be solved if we have a machine learning algorithm which can judge these students and score them accordingly. Another problem in bulk hiring is the management of time, which is solved by automating this process.

Many companies are working in this field, and have successfully achieved the near to human level accuracy, minimizing the common human errors. First, is the Gecko.AI which is an AI-based interview platform that works on AI, Sentiment Analysis and facial recognition. The beauty of this video-based evaluation bot is that it can conduct both live and offline interviews. The questions for interview are being set by recruiters that can later be played back for detailed analysis and review. The platform's AI-powered sentiment analysis engine scans each interview to provide deep insights on candidate attitude, positivity, and overall sentiment. The bot is designed in such a way that it even follows up with prospective candidates for interviews via email. And it continues to do that until candidate response [2]. Another company is MYA, which is a conversational AI assistant helps hiring teams of firms to form trust and confidence with candidates through open-ended, natural, and dynamic conversations. This AI-powered interview platform uses deep learning to deliver a human-like conversation to the candidate. Using semantic parsing, named entity recognition, and multiple intent classifications, Mya captures meaningful information from the candidate. The bot is designed in such a way that it understands the context, complex, multi-part statement, changed answers, or interjections and it can also change conversation direction. That is not all, Mya is in continuous learning the process as its machine learning algorithm with every conversation. The third is the Autoview by Aspiring Minds, which is an AI-powered interview bot. The platform uses video analytics, natural language processing, Machine Learning, and Speech recognition to carry out the interview process. It screens the candidates based on 4 parameters; based on candidate's facial expressions and gestures; sentiment analysis of voice and text; ability and knowledge required for the job role; and workplace competencies, cultural fitment, and personality. That is not all, the platform is available anytime so the students can take the interview as per his/her convenience. Another one is PANNA, which is a data-driven AI video interview platform that provides artificially intelligent hiring, an ever-growing repository of dynamic questions, expert evaluation, recorded interviewing, video conferencing and voice and face recognition. One of the best thing about this platform is that it blocks out proxy interviews and under-qualified applicants from large pools of candidates. The last one is Hirevue, which is the most popular in this genre of automation. It uses a combination of proprietary voice recognition software and licensed facial recognition software. With the help of an algorithm, the platform determines which candidates resemble the ideal candidate. And it does it all by analyzing traits such as body language, tone, and keywords used during the interview. One the algorithm is done with its job, it lets the recruiter know how much with its job, and it lets the recruiter know which candidates are at the top of the heap.

II. RELATED WORKS

Two main problems faced in the manual hiring processes are the time management and human biases, which lead to partial decisions and hence do not create a standard in the interview process. Let us consider the latter as of now. There are several types of biases that add up I a hiring decision, one example is related to the first impression. So when hiring recruiters to judge the candidate on the basis of first assessment of the interview and take the rest of the interview as a support for their initial analysis, this is also known as the confirmation bias [3]. "You don't get a second chance to make a first impression", based on this famous saying ChaLarn challenge introduced its First Impression challenge in 2016. The main aim of the competition was to evaluate the personality traits from video dataset consisting of 10,000 clips from YouTube videos which were made publicly available. The teams worked on the problem and provided analytics using machine learning and deep learning approaches, most appropriate result was given by NJU-LAMDA team which used separate models for audio and video processing. For the video analysis, they used Descriptor Aggregation Networks given by Wei *et al.* [4] and a pre-trained VGG-face model by Parkhi *et al.* [5] is used. A proposed by Ponce-López V. *et al.* [6], for audio, log filter bank features are used and a single fully connected layer with sigmoid activations are used. Florentine *et al.* [7] devised and also mentioned in the Introduction, companies have started using AI algorithms to automate the hiring processes which are believed to minimize human bias.

Courgeon *et al.* [8] built MACH -My Automated Conversation Coach, a software by MIT Media Lab that analyses the behavior, conversation, and overall personality to provide them feedback in order to help them improve their conversational skills. The software uses a virtual agent which response according to the user's expressions and responses. It asks the candidates a few questions, and after completing the interview gives them personalized feedback. On broad terms, it uses three features - Facial expressions, Prosody analysis, and Speech recognition. For male candidates, the agent is a male counselor and for female, it is female counselors. The gender- bias is given after a keen understanding of the sentiments of the candidates through their feedback responses. For the data analysis, two Facial Action Coding System (FACS) is used which analyze the nonverbal behavior, which is response pattern and average smile duration. After the analysis, the features that were observed were, expressiveness, listening behavior, acknowledgments, and duration of the interview. For creating a responsive interview bot and give the interview a more realistic experience, three main components were targeted, the appearance of the virtual agent, an interaction between the coach and the user, and effective feedback response. Two types of feedbacks are given to the candidates after the interview, one is Summary Feedback and the other is Focused feedback. The upper half of the feedback system includes only the smile based feature extraction and the lower half includes fours features namely, Total pause duration, Speaking rate, Weak language, and Pitch Variation. At the micro level, the Facial expressions were processed using Shore framework and AdaBoost algorithm. Smiles detection was evaluated on Cohn-Lanade dataset and Jaffe dataset. Head nods and shakes were detected using the "between eyes" region. Prosody Analysis uses pauses, loudness and pitch variations as the features to analyze the verbal part of the interview which was done using the open source speech processing toolkit Praat. Arm and posture animation were detected using motion capture. Lip synchronization was detected by phonemes which were generated using Cereproc. Other features included, Gaze behavior, facial animation, head orientation, timing, and synchronization. The analysis was done in three scenarios, first when the participants were given a set of educational videos to watch and interviewed afterward, second was allowed to practice on the MACH and watch their video, and the third group practiced on MACH, watched themselves on the video and also got their automated feedback from MACH. Counselor's rearing showed that the ratings for the third group were much higher than the first and second group, which showed that MACH was really solving the problem of automating the interviews.

The MACH feedback system used feature extraction to the analysis, so to make it analytically more strong, the introduction of machine learning techniques were done when I. Naim *et al.* [9] provided a framework to automate the interview process. A model was built using 138 recorded interviews which form up duration of 10.5 hours in total, of 69 internship searching students from Massachusetts Institute of Technology (MIT). Broadly, the machine learning model as it's the first step uses feature extraction and for extracting the features, this paper focuses on a few important features including facial expressions, language, and prosodic information. Ground truth values are derived after averaging over the ratings of 9 independent judges using Amazon Mechanical Turk service. The model predicts the value for friendliness, engagement, and excitement with the correlation score of more than 0.73. The proposed framework also emphasizes on the factors which are usually neglected by the candidates like language fluency, usage of fewer filler words, vocabulary, and smile. For result validation, the correlation is calculated between the ground truth and predictions, and Support Vector Regressor and Lasso Regressor are used for the building the model. Result shows, for almost all the features, the Support Vector Regressor(SVR) outperforms the Lasso Regressor.

Based on the MIT Dataset, I. Naim *et al.* [10] came up with a more detailed way of analyzing the interview performance using machine learning approaches and more concentrated feature extraction methodologies. MIT dataset developed during the research is publicly available and can be accessed by emailing the concerned authorities. Prediction framework uses two regression algorithms SVR and Lasso for getting the results based on the features extracted. Feature extraction includes prosodic, lexical, and facial features. After extraction, the features are normalized so that the regression algorithm is not biased towards any one particular feature and treats all the features equally. Prosodic features include speaking style, the rhythm and speech intonation. Social intent modeling is seen to be the application of prosodic features. V. Soman *et al.* [11] gave mechanism for prosodic features so that they are first analyzed over a particular answer, then averaged over all the other answers. To perform the prosody analysis, an open-source speech analysis tool PRAAT given by P. Boersma *et al.* [12] is used. Major prosodic features extracted include, information of pitch, vocal intensities, and spectral energy. Additionally, other prosodic features were also extracted which includes, pause duration, the percentage of silent frames, jitter, shimmer, breaks in speech. Lexical features provide details about the content of the interview and personality of the interviewee and are analyzed using the tool LIWC J. provided by W. Pennebaker *et al.* [13]. After that, a Greedy Backward Elimination feature selection researched by R. Caruana *et al.* [14] is applied which iterates over all the features and remove the features whose deletion doesn't have a significant impact on the score. Also, Latent Dirichlet Allocation (LDA) by D. M. Blei *et al.* [15] is used to learn the topics of the interview, by training it on a topic modeling dataset. For facial features, first faces are detected using Shore framework by B. Froba *et al.* [16], a binary classifier trained using the AdaBoost algorithm separates smiling faces from the neutral ones. Smile intensities, head gestures such as nods and shakes, and other features are extracted using a Constrained Local Model (CLM) based face tracker by J. M. Saragih *et al.* [17]. Additionally, three head pose features, roll, pitch, yaw, and are calculated using the Rotation matrix. For prediction, SVR regressor outperformed every other regressor and that too with linear kernel, the accuracy was better than non-linear kernels. Modern techniques like Long Short term

memory - Recurrent Neural Network (LSTM-RNN) by H. Sak *et al.* [18] architectures for speech recognition and detection of disfluency proposed by V. Zayats *et al.* [19] can lead to more than 85 % accuracy.

Communication do matters based on the scenario of the interview, the paper targets the difference between two cases, first one being the simultaneous video interview (no person is involved) and second being the face to face interviews (when the interviewer is involved) as studied by Rasipuram *et al.* [20]. The dataset formed, contains 106 interviews in both the settings and Observations show that the behavioral perception is slightly different in both the cases. Feature extraction concludes low-level features based on audio, visual and lexical features and the prediction algorithm used is SVR and Logistic Regression. As per the analysis in the two modes, the confusion is that 45% of the participants had a better rate of speech when in a face-to-face setting, and 40% were more expressive in video interface mode. A possible reason for better expressiveness in the video interface mode might be that candidates might not have suppressed expressions. Similarly, expressions were relevant for face-to-face and enthusiasm was relevant for a video interview. In a face-to-face interview, the lexical features were more dominant and for a video interview, the analysis shows that prosodic features become more vital than others.

Instead of analyzing the interview as an audio-visual problem, specialized approaches have also come into the limelight like fluency prediction in interface based interviews as given by S. Rasipuram *et al.* [21]. Fluency is known to be one of the most important factors in an interview which can deviate the result drastically. Fluency is predicted by extracting the lexical, prosodic and speaking based features. After feature extraction, normalization and fine-tuning, supervised classification methods are applied for prediction. Correlation analyses show the rate of speech as the most important and highly correlated attribute for fluency. Also, pitch and high energy indicate the level of confidence which affects the fluency.

A vital part of the interview is the assessment of communication skills, which helps in determining the personality and character of the interviewee by Rao *et al.* [22]. Based on the features extracted and algorithm trained, the proposed model achieves 75% accuracy. Three different scenarios which are considered are Video interview, written interview, and a short essay. Dataset consists of 100 interviews in the three given settings and then a deep analysis is done. Features important for communication skills come out to be eye contact, writing and speaking fluency and grammar. Comparative analysis shows that video and written interview are only different as per the mode of communication. The written interview and short essay are different from the time dimension and the question type. 71% of candidates who performed miserably in spoken communication also performed badly in written communication. This showed that the performance of the candidate is independent of the mode of communication.

CONCLUSION

There has been a lot of research in the field of automatic prediction of job interviews in the last decade. There have been multifarious approaches towards the audio-visual regression prediction of scoring the interviews. The main conclusion of this paper is that there has been researching done in the past but the maximum accuracy achieved is 85% which can be improved by using the state of the art deep learning methodologies as well as by fine-tuning the features. This paper tries to find the research gap in the field of automatic job prediction to tries to build a platform for further research work.

REFERENCES

[1] https://www.twinemployment.com/blog/8-surprising-statistics-about-interviews

[2] https://indianceo.in/business/ai-powered-hr-systems-indian-scenario/

[3] Wired. 2015. Here's Google's secret to hiring the best people. https://www.wired.com/2015/04/hire-like- google/.

[4] Wei, X.S., Luo, J.H., Wu, J.: Selective convolutional descriptor aggregation for fine-grained image retrieval. arXiv (2016)

[5] Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC. Volume 1. (2015)

[6] Ponce-López V. et al. (2016) ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results. In: Hua G., Jégou H. (eds) Computer Vision – ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science, vol 9915. Springer, Cham

[7] Florentine, S. 2016. How artificial intelligence can eliminate bias in hiring. https://www.cio.com/article/3152798/artificial-intelligence/how-artificial-intelligence-can-eliminate-bias- in-hiring.html

[8] Courgeon, Matthieu & Martin, Jean-Claude & Mutlu, Bilge & Picard, Rosalind & Ehasanul Hoque, Mohammed. (2014). MACH: My automated conversation coach. UbiComp 2013 - Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 10.1145/2493432.2493502.

[9] 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," 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, 2015, pp. 1-6.

[10] I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque, "Automated Analysis and Prediction of Job Interview Performance," in IEEE Transactions on Affective Computing, vol. 9, no. 2, pp. 191-204, 1 April-June 2018.

[11] V. Soman and A. Madan, "Social signaling: Predicting the outcome of job interviews from vocal tone and prosody," in ICASSP. Dallas, Texas, USA: IEEE, 2010.

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1. 05)[computer program]. retrieved may 1, 2009," 2009.

[13] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic Inquiry and Word Count: LIWC 2001," Mahwah: Lawrence Erlbaum Associates, vol. 71, 2001.

[**14**] R. Caruana and D. Freitag, "Greedy attribute selection." in International Conference on Machine Learning (ICML), 1994, pp. 28–36.

[**15**] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," The Journal of Machine Learning Research, vol. 3, pp.

993–1022, 2003.

[16] B. Froba and A. Ernst, "Face detection with the modified census transform," in Automatic Face and Gesture Recognition (FG). IEEE, 2004, pp. 91–96.

[17] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009, pp. 1034–1041.

[18] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[19] V. Zayats, M. Ostendorf, and H. Hajishirzi, "Disfluency detection using a bidirectional lstm," arXiv preprint arXiv:1604.03209, 2016.

[20] Rasipuram, Sowmya & Rao, Pooja & Jayagopi, Dinesh. (2016). Asynchronous video interviews vs. face-to-face interviews for communication skill measurement: a systematic study. 370-377. 10.1145/2993148.2993183.

[21] S. Rasipuram, S. B. P. Rao and D. B. Jayagopi, "Automatic prediction of fluency in interface-based interviews," 2016 IEEE Annual India Conference (INDICON), Bangalore, 2016, pp. 1-6.

[22] Rao, Pooja & Rasipuram, Sowmya & Das, Rahul & Jayagopi, Dinesh. (2017). Automatic assessment of communication skill in non-conventional interview settings: a comparative study. 221-229. 10.1145/3136755.3136756.