Interview Score Prediction and Analysis Using Audio Features and NLP

Aditya Jha
(Department of CSE)
NIT Calicut
Calicut-673601, India
aditya_b180648cs@nitc.ac.in

2 Ritik Gautam
Department of CSE
NIT Calicut
Calicut-673601, India
ritik_b180630cs@nitc.ac.in

3 Palash Bajpai
Department of CSE
National Institute of Technology Calicut
Calicut-673601, India
palash_b180759cs@nitc.ac.in

Abstract—**The traditional interview is a test of personality traits for many hiring managers. However, there are several limitations to traditional methods of interviewing. In addition to taking up time, it can be difficult for some candidates or interviewers to get a read on the candidate because there's no way to know what their voice sounds like under stress or how to accurately score candidates based on the way they answer the questions. Prosodic features can help to get insights on emotions, confidence, anxiety, etc which can help interviewers to judge the candidate. Also, by extracting the transcript from available audio and applying NLP and other machine learning models we can score the answers given by interviewees. Thus, our project uses machine learning to facilitate the faster assessment of candidates by using prosodic features and NLP, ultimately improving the interviewer's ability to make hiring decisions based on more reliable data that is less susceptible to human biases**

Index Terms—**Nonverbal behavior prediction, Job Interviews, Lasso, Deep Learning, Multi-Modal Approach**

I. Introduction

In recent years, online video-based interviews have been increasingly used in the hiring process. For example, HireVue, a major vendor in the online video interview hosting market, has reportedly provided its service to many Fortune 500 companies. Conducting online video-based interviews brings many benefits to both interviewers and interviewees, including the convenience of offline 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. Interviewees must adapt their multimodal behaviors, such as speech content, prosody, and nonverbal or facial cues to effectively communicate their qualifications in a limited amount of time. The success or failure of the interviewee's effort is traditionally evaluated by the interviewer, either through personality or quantitative ratings or both according to the need of the company. We can easily interpret the meaning of our verbal and nonverbal behavior during face-to-face interactions. However, we often cannot quantify how the combination of these behaviors affects our interpersonal communications. An emerging alternative to the traditional human-only interview assessment model is to augment human judgment with an automated assessment of interview performance. The style of speaking, prosody, facial expression, and language reflect valuable information about one's personality and mental state. Understanding the relative influence of these individual features and their dependence can provide crucial insight regarding job interviews. These features include facial expressions (e.g., smiles, head gestures, facial tracking points), language (e.g., word counts), and prosodic information (e.g., pitch, intonation, and pauses) of the interviewees. We are proposing a model to design and implement an automated prediction framework for quantifying the ratings of job interviews based on audio features (prosodic, lexical), given the audio recordings. The prediction framework automatically extracts a diverse set of multimodal features (lexical and prosodic) and quantifies the overall interview performance, the likelihood of getting hired, and 14 other social traits relevant to the job interview process. We would be extending the idea of interview score prediction to more like a virtual coach or a tool that would help the candidates to improve them by going through the feedback and suggestions given by the tool. It would be a kind of interactive tool that would synthesize the actual features that our computational framework would need and try giving them suggestions, summary feedback, and detailed feedback on the specific areas that the candidate wants he can go through the feedback and try to improve based on suggestions. We are planning to use prosodic and lexical features to build our framework focusing more on the questionnaire section and using NLP and deep learning to increase the accuracy for the questionnaire section and also it would help the candidates to get a better understanding of the interviews.

II. Problem Definition

To build a suggestion-based interview analysis tool that can help companies to score candidates based on lexical and prosodic features also gives feedback and suggestions to interviewees.

III. Background and Domain Details

In this section, we discuss existing relevant work on nonverbal behavior prediction using automatically extracted features. We particularly focus on the social cues that are relevant to job interviews and face-to-face interactions. As the earlier study showed that motion cues and gesture dynamics plays important role in determining the emotions later this study was further explored and found to be convincing and more research was done on this area. Later many research was conducted and it was found to be very convincing for determining one's emotions. These emotions were then proved to use for the prediction of the candidate's performance during the job interview. Later many frameworks were developed incorporating many features to predict job interview performance such as audio emotions, facial emotions, etc. This domain particularly involves signal processing, facial coordinates, Action Units used for mapping facial coordinates to emotions. Openface is one such prebuilt tool available. Similarly for audio processing, many prebuilt tools and frameworks are available one of them is the shore framework. Audio features include many features such as spectral features, chroma features, pitch, etc. In addition to that Natural Language Processing can be used to handle the lexical features in contributions to deep learning and machine learning.

IV. LITERATURE SURVEY

- [1] Addresses the challenge of automated understanding of multimodal human interactions, including facial expression, prosody, and language. They used the open-source speech analysis tool PRAAT for prosody analysis, LIWC for lexical features. SVR and LASSO were used as regression models. Ratings predicted by the model are based on social and behavioral skills only.
- [2] Proposed a machine learning-based method to check a candidate's aptitude and personality score based on uploaded CV. The TF-IDF algorithm is used to perform the analysis as a graph in terms of the programming skills on x-axis and the respective scores on y-axis.
- [3] It proposes Social Signal Processing (SSP) which provides a general framework of using multimodal sensing and machine perception to analyze human communication including job interviews. BARS rating method was also used. FE method that is highlighted by applying the Neural Network (NN) based doc2vec paradigm to obtain effective visual features.
- [4] SER has also been used by the Convolutional Neural Network (CNN) and CNN Alex Net models. (FACS)Facial Action Coding System use for facial recognition which assigns a numerical value to each facial moment. DNN Model was used for resume parsing and verification.
- [5] Addresses the challenge of HPC- High-performing clusters for prediction of job logs use regression and neural networks models to manage the prediction of new jobs in the job logs so that performance of HPC can be increased.
- [6] This is an automated coach tool developed to provide the interviewee with some insights about the interview. This focuses on three features, expressiveness, response pattern, and acknowledgments. The virtual coach was designed by keeping in mind facial, non-behavioral synthesis Animation of the coach, etc.
- [7] A software tool for real-time fully automated coding of facial expression. It provides estimates of facial action unit intensities for 19 AUs from the Facial Action Unit Coding System (FACS) a, as well as probability estimates for the 6 prototypical emotions (happiness, sadness, surprise, anger, disgust, and fear).
- [8] This also uses nonverbal cues to make a computational framework for hire ability. The basic approach is extracting the video cues and the audio cues to build a baseline model using ridge regression and then using questionnaire data to measure correlation with hire ability score. The R2 value validated the model up to 36 percent.
- [9]The system consists of a speech recognizer trained on non-native English speech data, a feature computation module, using speech recognizer output to compute a set of mostly fluency based features, and multiple regression scoring models which predict a speaking proficiency score for every test item response, using a subset of the features generated by the previous component.
- [10]This paper quantifies the non-linguistic speaking style of engineering school students in practice job inter- views, using features extracted from their vocal tone and prosody. It finds that successful candidates have a characteristic speaking style and these vocal features can be used to build a predictive model of the interview outcomes, with over 85 percent accuracy

V. DATASET

We used the MIT Interview Dataset, which consists of 138 audio-visual recordings of mock interviews with internship-seeking students from the Massachusetts Institute of Technology (MIT). The total duration of our interview videos is nearly 10.5 hours (on average, 4.7 minutes per interview, for 138 interview videos). Initially, 90 MIT juniors participated in the mock interviews. All participants were native English speakers. The interviews were conducted by two professional MIT career counselors who had over five years of experience. For each participant, two rounds of mock interviews were conducted: before and after interview intervention. During each interview session, the counselor asked interviewees five different questions, which were recommended by the MIT Career Services. These five questions were presented in the following order by the counselors to the participants:

- So please tell me about yourself.
- Tell me about a time when you demonstrated leadership.
- Tell me about a time when you were working with a team and faced a challenge. How did you overcome the problem?
- What is one of your weaknesses and how do you plan to overcome it?
- Now, why do you think we should hire you?

VI. DESIGN AND INPUT-OUTPUT IDENTIFICATION

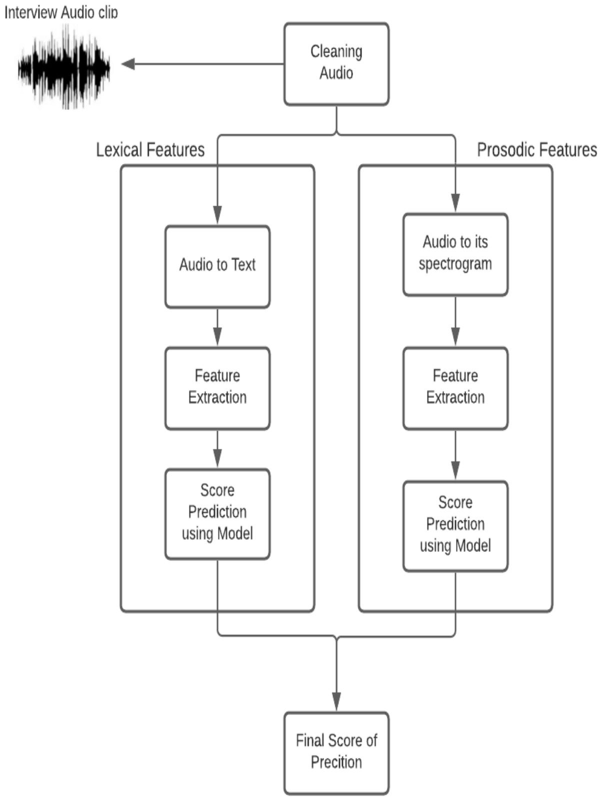

Fig. 1. Implementation approach.

A. Input:

For training purposes, our input to the model will be audio files for the audio analysis model and we will generate transcripts from these audio files which will be used by the lexical analysis model for training purposes. For the working tool, we can either have a video from which we will extract the audio or audio file itself.

B. Output:

It would be a score given to each candidate out of 100. We will also try to make some feedback system so that it helps candidates to know things they should work on and improve for the next interviews. In this way, this tool will help in both interview analysis and interview preparation or practice tool.

C. Framework explanation:

We are trying to build each component independently and will try to build a framework that can combine outputs from both lexical and audio features to generate a score out of 10. This will help us to follow a modular approach. So we would be building two components one for prosodic and one for lexical. Both the components would follow the same general approach converting the data into actual data that we can understand and visualize extracting the features and building a model that would give a score to that candidate out of 100 for now and then merging both the scores based on certain criteria to generate the final score. Also try to give suggestions by analyzing what factors affecting his interview positively and negatively, so to help candidates to analyze their interviews and increase trust in our model.

D. Final Tool:

A suggestion-based interview analysis tool that can help companies to score candidates on basis of lexical and prosodic features also gives feedback and suggestions to the interviewee. As of now, we are working on only HR interviews however this can be extended to all types of interviews.

VII. METHODOLOGY

A. Step 1: Feature Extraction

Both audio and lexical features contain a large number of features that may affect the interview scores. We have extracted quite a large number of features and we will use them to make our model. Till now we have extracted features for audio and transcript as a whole, but we will extract features for fragmented audio and transcript files so that we will have a larger dataset for training our models.

Lexical analysis: For lexical analysis, we have extracted features like count of adverbs, verbs, words representing sadness, words expressing sorrow, etc. We currently have 23 features for lexical analysis which cover all factors affecting emotions, confidence, and other personality traits. Since we have scores for these interviews we can train the model to find which types of words affect the score of interviews and by how much.

Audio analysis: From audio files, we have extracted features that affect tone, pitch, energy, and other related features which affect the audio. We have used pyAudio analysis library to extract the features from .wav audio files. We currently have 204 features extracted however we will need to select the most relevant features amongst them.

B. Step 2: Data Preprocessing

We have fragmented audio files into subfiles for generating more training data for our models. We will have to generate separate transcripts for all. Also till now, we have extracted features for the whole audio and transcripts but later we will extract features for individual sub-audio files. We will consider the same scores for sub-audio files as of their parent audio files however the values for the features will change since for lexical count of words will change and for audio files also factors will be different than parent audios.

C. Step 3: Machine Learning Model

Since we have 2 separate components for our project hence we will make 2 separate machine learning models for both. Lexical model will train on features we extracted using natural language processing which is word counts from transcript, for audio analysis model input will be the factors we extracted using pyAudio analysis library, which contains audio related factors like tone, pitch, etc. These factors affect personality and human bahaviour a lot and hence are important for analyzing interview scores.

D. Step 4: Output

Till now we will have outputs from both lexical and audio analysis models, however, we don't know which factors affect the final score and by how much. So we will try to build a standard framework that will incorporate all factors and try to evaluate the interview of the candidate. Since we will also know about how different factors are affecting the score, we can develop a feedback mechanism that will help the interviewee to know his/her weak and strong points so that they can improve their skills for the next interview.

VIII. WORK DONE

- We successfully collected data from MIT.
- We have understood the problem and the requirements to solve the problem efficiently and more accurately.
- We have fragmented our audio files into subparts based on the duration of questions asked, each part contains one question and its answer, and this helps to generate more data, which will help our model to work more accurately.
- We have extracted features from both audio files and stored their values in excel sheets.
- We are trying to modularize our approach.
- We have made a basic model for lexical analysis using a multi-output regression model, however, we will try more models to get better accuracy.
- For audio analysis also we have used a multi-output regression model using all the features, for now, we will try more models after selecting the most relevant features.

IX. Model

A. Lexical Analysis

Input dataset: We used Natural Language Toolkit (nltk) library to find the count of filler words. NLTK consists of the most common algorithms such as tokenizing, sentiment analysis, word extraction, etc which will help to analyze and preprocess any text.

Working model: We have used a multi-output regression algorithm with k-fold validation. Since from the inputs we need to find values of many different factors like recommend hiring, engagement, friendliness, etc, thus a multi-output regression model is used. K-fold validation is used to validate the model on go while training.

Output: We have used a multi-output model hence we will get values for all different factors which we are trying to use to predict the score at the end.

B. Audio Analysis

Input dataset: We have used pyAudio analysis library which is used for audio feature extraction, analysis, classification, and related applications. We used our .wav files as input and got values for different features which in turn affects the quality of voice, pitch, energy etc. The meaning and value of these features are hard to understand and need some study. We will have to find what factors of audio are affected by which feature to select the best features for audio analysis.

Working model: We have used the same multi-output regression algorithm with k-fold validation however with a different input set. Since we have common values for both audio and transcript files, from the inputs we need to find values of many different factors like recommend hiring, engagement, friendliness, etc, thus a multi-output regression model is used. K-fold validation is used to validate the model on go while training. This model works independently of the lexical analysis model. We will further try different other models to use for the given purpose.

Output: We will get values for all different factors which we are trying to use to predict the score through audio of a person.

X. WORK PLAN

- Select the features or stimulate the features from the input select and select the most reliable features from the input.
- Modify the dataset to include data for sub-audio files and transcripts and include only the most contributing factors by doing using feature selection techniques.
- Try different models for lexical analysis to know which one works best.
- Build a model for audio analysis by using the most reliable features.
- Build a framework to score interviews based on outputs from both lexical and audio analysis models.
- Modularize the approach and try to build components accordingly so that we can parallelize the work and do the necessary changes so updating one component will not disturb the rest of the components.
- Test each component separately. The components should be loosely coupled and split the features for each component eg : (One component for prosodic features, one for lexical, etc).
- Work on the accuracy of all models and test them on different

test cases.

- Try to build a suggestion-based tool that takes our model and gives feedback and suggestions to the candidate.
- Present the final report and give a demo of our work to the panel.

XI. SUMMARY

The scope of this project has not been explored much, and many new ideas can be implemented. Interview analysis is used by HR and companies to determine the suitability of a person for the job. This project can even be used by students to train and prepare themselves for interviews. With both audio and lexical features, we use two important techniques of human behavior analysis. This project also includes the use of quite different areas of machine learning and thus we will be learning quite a new thing. I hope this project gets built according to company standards and we can use this project in real life.

REFERENCES

[1]I. Naim, D. Gildea, Md. I.Tanveer and Md. E. Hoque, “Automated Analysis and Prediction of Job Interview Performance” - IEEE Transactions on affective computing, VOL. 9, No. 2, April-June 2018

[2] Jayashree Rout, Sudhir Bagade, Pooja Yede, and Nirmiti Patil, “Personality Evaluation and CV Analysis using Machine Learning Algorithm”
- IJCSE Vol.-7 E-ISSN: 2347-2693 Issue-5, May 2019

[3] Lei Chen, Gary Feng, Chee Wee Leong, Blair Lehman, Michelle Martin-Raugh, Harrison Kell, Chong Min Lee, and Su-Youn Yoon, “Automated Scoring of Interview Videos using Doc2Vec Multimodal Feature Extraction Paradigm” - ICMI ’16: Proceedings of the 18th ACM International Conference on Multimodal InteractionOctober 2016

[4] Supriya Anand, Nihar Gupta, Mayesh Mulay, and Abhimanyu Sherawat “Personality Recognition and Video Interview Analysis” - International Journal of Engineering Research and Technology (IJERT), ISSN: 2278-0181, IJERTV10IS050122 Vol. 10, Issue 05, May-2021

[5] Zhengxiong Hou, Shuxin Zhao, Chao Yin, Yunlan Wang, Jianhua Gu, and Xingshe Zhou “Machine Learning-based Performance Analysis and Prediction of Jobs on a HPC Cluster” - in IEEE Access, INSPEC Accession Number: 19452885, DOI: 10.1109/PDCAT46702.2019.00053, 12 March 2020

[6] ZGwen Littlewort, Jacob Whitehill , Tingfan Wu , Ian Fasel , Mark Frank , Javier Movellan , and Marian Bartlett “The Computer Expression Recognition Toolbox (CERT)” - in IEEE Access, INSPEC Accession Number: 12007742, DOI: 10.1109/FG.2011.5771414, 19 May 2011

[7] Laurent Son Nguyen, Denise Frauendorfer, Marianne Schmid Mast, and Daniel Gatica-Perez, “Hire me: Computational Inference of Hirability in Employment Interviews Based on Nonverbal Behavior” - in IEEE Transactions on Multimedia, Vol. 16, No. 4, June 2014.

- Matthieu Courgeon, Jean Claude Martin, Bilge Mutlu, and Mohammed Ehasanul Hoque, “MACH: My automated conversation coach” - DOI: 10.1145/2493432.2493502, issue: Dec 2.

[8] Zechner, Klaus Higgins, Derrick and Xi, Xiaoming and Williamson, David. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. Speech Communication. 51. 883-895. 10.1016/j.specom.2009.04.009.

[9] Vikrant Soman, Anmol Madan Electrical Engineering Department MIT Media Laboratory University of Wisconsin-Madison Massachusetts Institute of Technology SOCIAL SIGNALING PREDICTING THE OUTCOME OF JOB INTERVIEWS FROM VOCAL TONE AND PROSODY Appears: IEEE In’tl Conference on Acoustics, Speech and Signal Processing, Dallas TX March 2009