Critical Speech-Analysis with Explanations

Ahmed Ewva
Technical University of Munich
ahmed.ewva@tum.de

Stefan Frisch
Technical University of Munich
stefan.frisch@tum.de

Ruiyun Xie
Technical University of Munich
ruiyun.xie@tum.de

Abstract

In a job interview setting, personality traits, behavioral cues and soft skills are evaluated alongside the experience and hard skills. It is therefore important for potential interviewees to be aware of his/her positive and negative attributes. This paper provides a multi-modal approach to assess and give feedback for an interviewee, based on his/her performance in an interview simulation. We extract prosodic, textual, and facial features and train machine learning models to predict relevant labels such as emotions and fluency and analyze the association between these labels and the quality of an interview. Explainability techniques such as LIME are used to highlight the most prominent features and thereby make the system more understandable for the end user

1 Introduction

Job interviews are an ubiquitous process where a recruiter evaluate the skills, knowledge, and abilities of a candidate to assess his/her suitability for an open position. Generally, a job interview is a face-to-face social interaction between two people. Therefore, many factors such as body language, facial expressions, intonation, verbal and nonverbal cues come into play to evaluate a candidate. For instance, studies in psychology have shown that maintaining eye contact, smiling, and using louder voice contribute positively to our interpersonal communication and thus have a good impact during an interview (Huffcutt et al., 2001). In this work, we present the **A**utomatic **F**eedback generation **F**ramework for job **I**nterviews **AFFI**. AFFI is a multi-modal data driven computational framework to provide automatic specific feedback to job interviewees.

We start by extracting different lexical, prosodic and facial features. Some of features are used to train machine learning models to provide labels that corresponds to verbal and nonverbal signals like emotions. While others are direct inputs to statistical analysis. The feedback is given by associating high-level features and labels on the one hand, with the performance scores of interviews on the other hand. For instance, our results in section 5 show a strong negative correlation between negative emotions and the performance score.

2 Dataset

The MIT dataset serves as the evaluation and training basis for the automated job interview feedback of our system (Naim et al., 2018). It consists of 138 mock job interviews by 69 MIT students. Each interview is on average 4.7 minutes long and all interviewees were asked the same 5 questions.

> Q1: So please tell me about yourself.
>
> Q2: Tell me about a time when you demonstrated leadership.
>
> Q3: Tell me about a time when you were working with a team and faced a challenge. How did you overcome the problem?
>
> Q4: What is one of your weaknesses and how do you plan to overcome it?
>
> Q5: Now, why do you think we should hire you?

The interviews have been transcribed and rated by Amazon Mechanical Turk workers along 16 categories on a scale from 1 to 7. The categories are: *Overall Rating, Recommend Hiring, Engagement, Excitement, Eye Contact, Smile, Friendliness, Speaking Rate, No Fillers*. In figure 1, an example of the camera setup in one of the mock interviews can be observed.

Figure 1: Camera and setup example for the MIT mock interviews (Naim et al., 2018).

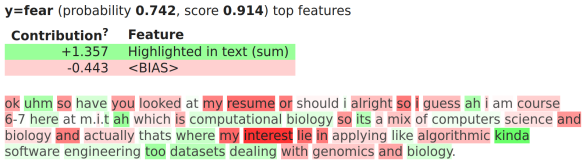

Figure 2: Example of a lime generated explanation for a sentiment score of one sentence in the interview pp89 from the MIT data set.

3 Approach

For the feedback, several different features are automatically extracted from each interview. There are three general categories of features: 3.1 lexical features, 3.2 prosodic features and 3.3 facial features. The analysis and feedback creation is also done separately for the three dimensions and only aggregated in a last step to make the full feedback generation process more tangible to the user. The emotion detection and analysis represents a major part in each of the three dimensions, because it was shown that emotions play a vital role in first impression on video blogs (Biel et al., 2012) and during job interviews (Naim et al., 2018).

3.1 Lexical analysis

The lexical analysis of the MIT interview transcripts is based on three different kinds of features:

1. Word count like features
2. **L**inguistic **I**nquiry **W**ord **C**ount (**LIWC**)
3. Sentiment analysis with **BERT**

Word count like features are included in the analysis since even though these are low-level features, they are still valuable insights in this context. As an example of why they might be important, please consider the following: Interviewees with a very high speaking rate, which can be measured in words per second, are perceived more nervous and less well informed compared to more measured candidates. Additionally, the number of unique words per second is also extracted. The usage of more unique words can be an indicator for how confident and knowledgeable an interviewee is in a certain domain.

The linguistic inquiry word count is a method based on word. It extracts over 90 psycholinguistic categories from a text (Pennebaker et al., 2015). These categories include psychological constructs (positive and negative emotions, anxiety, anger), linguistic dimensions (pronouns, verbs, nouns) and informal language markers (filler words, non-fluencies, netspeaking). In a similar paper where the automated interview performance assessment has been studied, 23 out of these 90 categories have been selected (Naim et al., 2018).

For the sentence level multi-class sentiment analysis with BERT it was unfortunately not possible to work with the labels provided by the MIT dataset. These sentiment scores are only given on the per interview basis and not on the sentence level which makes it unsuitable for fine-tuning BERT (Devlin et al., 2019). Instead, a balanced dataset was created from the dailydialog (Li et al., 2017), emotion-stimulus (Ghazi et al., 2015) and isear (Scherer and Wallbott, 1994) datasets. This new dataset is comprised of about 11000 sentences with the labels: joy, sadness, anger, neutral and fear. The performance of BERT on the test set is 83%. The performance on the MIT dataset was manually inspected. In addition to the pure sentiment scores, the user is provided with an explanation for each sentence. This should help the interviewee understand the output of the sentiment analysis. The explanations are generated with LIME (Ribeiro et al., 2016). An example of one explanation can be seen in figure 2. Explanations for all sentences are also added to the 6 dashboard.

3.2 Prosodic analysis

Prosodic features are important for characterizing the speaking style, the rhythm and intonation of speech which are an important factor that impacts the quality of an interview. The prosodic analysis consists of the following two approches:

1. Low-level feature extraction and analysis with pretrained models.
2. Direct extraction of high-level features and analysis of their impact on the interview.

3.2.1 Low-level feature extraction and models training

As a first step the interviewer is separated from the interviewee with and unsupervised clustering algorithm. Then the interviewee's audio is divided into small chunks of 3 seconds each. Then we use the libraries: PyAudioAnalysis (Giannakopoulos, 2015) and PRAAT (Boersma and Weenink, 2021) to extract 150 low-level features from each audio segment. These low-level prosodic features are based on frequency, intensity, tone, zero-crossing rate, energy, jitter etc. After extracting these features, a sentiment classification model and a fluency classification model are applied for each segment and the results are aggregated for the entire interview.

To obtain the sentiments and the fluency models, we relied on two other datasets.

- Avalinguo audio dataset (Preciado-Grijalva and Brena, 2018): contains 1424 audio, classified into 3 classes (high, intermediate, and low)
- RAVDESS (Livingstone and Russo, 2018), TESS (Dupuis and Pichora-Fuller, 2010), SAVEE (Jackson and Haq, 2014), CREMA-D (Cao et al., 2014) datasets : contain audio files classified into 8 emotions: angry, sad, disgust, surprise, calm, neutral, happy, and sad (calm and neural are grouped together in the following).

For both tasks we extracted the same 150 features and trained multiple classification models. For the sentiment classification the best results were reached with an SVM model with an accuracy of 70%. For the fluency classification the best results were reached with an SVM model with an accuracy of 88%.

After training and testing the models, explanations are generated using LIME. For each singular prediction, the main features that lead to that particular prediction are displayed 6 alongside their weights.

3.2.2 High-level feature extraction and analysis

For the second approach we use the libraries PRAAT (Boersma and Weenink, 2021) and Myprosody [1] to extract the following features:

- Number of syllables
- Number of pauses
- Speaking duration
- Rate of speech: number of syllable over the total duration
- Articulation rate: number of syllables over the speaking duration
- Balance: speaking duration over the total duration

These features provide valuable information regarding the speaking styles of the interviewees especially when the comparison to good performing interviews is illustrated.

3.3 Facial analysis

The facial expressions are usually the most important factor in the first impression and how others perceive one's appearance. Most importantly, one can infer the emotion of the person directly from the facial expressions. For the facial analysis, we use the video files from the MIT data set. The general approach for processing the video files is to extract one frame per second, crop and extract the faces and perform three different tasks on the resulting images:

- Facial action units (FAU) extraction
- Facial emotion recognition
- Valence and arousal level estimation

The facial action units are components of the Facial Action Coding system (Ekman and Friesen, 1978), which is an anatomically based system for describing all visually discernible facial movement. The extraction was done with OpenFace 2.0 (Baltrusaitis et al., 2018), which is a powerful tool for facial analysis. It can detect faces in an image or a video and their landmarks, estimate head poses and extract 18 facial action units as shown in Figure 3. The intensity and presence of all AUs are predicted, except for AU28 only the presence is detected.

For the emotion recognition, a pre-trained model based on a CNN architecture[2] is used to extract

[1] https://github.com/Shahabks/myprosody
[2] https://github.com/justinshenk/fer

AU	Full name	Illustration
AU1	INNER BROW RAISER	
AU2	OUTER BROW RAISER	
AU4	BROW LOWERER	
AU5	UPPER LID RAISER	
AU6	CHEEK RAISER	
AU7	LID TIGHTENER	
AU9	NOSE WRINKLER	
AU10	UPPER LIP RAISER	
AU12	LIP CORNER PULLER	
AU14	DIMPLER	
AU15	LIP CORNER DEPRESSOR	
AU17	CHIN RAISER	
AU20	LIP STRETCHED	
AU23	LIP TIGHTENER	
AU25	LIPS PART	
AU26	JAW DROP	
AU28	LIP SUCK	
AU45	BLINK	

Figure 3: List of available facial action units in OpenFace 2.0 (Baltrusaitis et al., 2018).

the possibility for 7 different emotions: angry, sad, disgust, surprise, neutral, happy, and fear.

After these two tasks, the relationship between the facial action units, the emotions and other labels from the MIT dataset is investigated. For that, statistical analysis is performed, which will be further elaborated in section 5.

Last by not least, the valence and arousal level of a person in the video are estimated. Valence describes how positive and negative an emotion is, while arousal describes how active or calm the person is. For the estimation, a pre-trained model with a CNN and RNN architecture (Deng et al., 2020) is applied, which returns two scores in the interval [-1, 1].

4 Timestamp creation

To combine the textual analysis with the audio and video it was necessary to create timestamps for each sentence. Specifically the time when the interviewee starts a sentence and stops it. No timestamps are created for the interviewer's sentences since his/her language, facial features and tone of voice are not analyzed. So the interviewer and interviewee are separated in the aforementioned way and then split on pauses to create small audio chunks that resemble sentences. These are translated with the Google Speech-to-Text API(goo). Finally, a matching between the human transcription and the google transcription is calculated with Cosine similarity and relative position in the text. The results for one example interview are depicted in figure 4. One can see that there is a matching counterpart for most sentences from the google output. However, for some very short sentences either on the MIT or the google side there are no reasonable matching partners. For these the timestamps are interpolated.

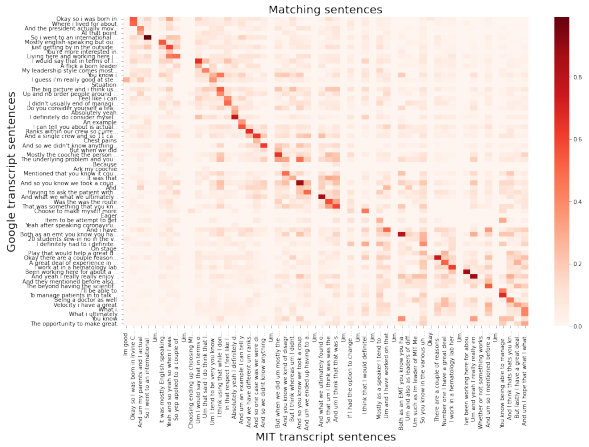

Figure 4: Timestamp example for interview pp98.

5 Results

After the feature extraction and timestamp creation, meaningful results from the features need to be drawn in order to provide feedback. Our approach is to apply the feature extraction methods from each domain on the 138 interviews from the MIT data set. We used a 'Recommended Hiring' score from the data set as the performance score and clustered the top 20% as good interviewees and the remaining 80% as bad interviewees. The threshold is used since it corresponds to the real world that most interviewees do not get hired for one job. Then the distribution of the features is plotted- one example for the facial emotion distribution is shown in figure 5 - where the blue area is for the bad interviewees and yellow area for the good interviewees. The mean and variance of the good interviewees' scores are also computed, which is shown as the blue lines in the figure 5 and used this interval later for the evaluation of a single interviewee's performance. An interviewee is considered performing poorly against the others if his/her score for one feature does not lie in this interval. Further statistical analysis such as correlation computation is also applied to gain more insights of the data.

For the lexical features, the analysis of the MIT

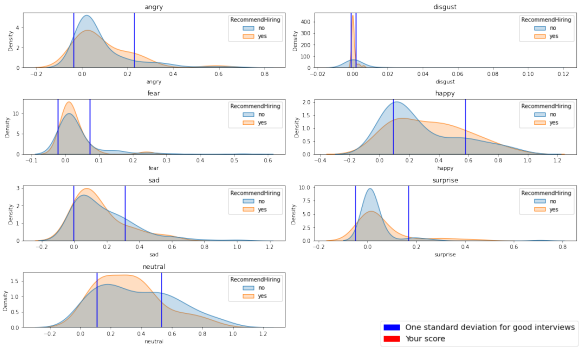

Figure 5: Distribution graph for the facial emotion.

interviews shows that a good interviewee uses more joyful, less fearful and slightly more angry sentences than an average/bad interviewee. A deeper analysis of the angry label revealed that these sentences are mostly situations with short distinct answers. The reason for this bias probably stems from the training dataset. Short very distinct answers in daily situation like in the dailydialog dataset usually point towards the angry emotion. Additionally, it was found that an average good interviewee speak with less filler words and more unique words.

On the prosodic level, the results show that the performance scores have strong negative correlation with the fear emotion and low fluencies, and strong positive correlation with the happiness and disgust emotions. However, the results do not show a significant statistical association between the scores and the others emotions. It is observed that when the number of pauses, speaking rate and articulation are too high or too low the overall score of the interview tends to decrease.

From the analysis of the facial features, a good interviewee is proven to show more happy and less neutral or sad emotion. In order to explain this observation, the relationship between the facial action units, the emotions and other labels we get from the MIT dataset is further investigated. The result shows, that some facial action units which are highly correlated to the happy emotion also correlate with positive labels like 'Smiled', 'Friendly' and 'Authentic'. This means an interviewee showing happy emotion will be perceived more positively with his or her facial movements, which also explains why they have a higher score. On the other hand, facial action units correlating with neutral emotion also correlate negatively with the label 'Authentic', explaining why an interviewee showing neutral emotion gets a lower score.

6 Dashboard

In order to integrate and display all extracted features and give the users valuable and understandable feedback based on our results from section 5, a dashboard[3] with the framework flask[4] is implemented.

On the first page of the dashboard, one can select and upload a video and will be then redirected to an overview page. The overview page is separated into 4 areas as shown in figure 6, one for the video and the other three for each of the domain, showing the most important features and explanations for the features. In particular, this means the highlighting of facial part responsible for the emotion and the results of the LIME explainer used for the lexical and prosodic features. The features are updated on the fly, i.e, the facial features are updated every second for each frame, and the lexical and prosodic features are updated with the timestamp generated as described in section 4. The area of each domain is linked to an additional page with more detailed feedback and there is also a link for general feedback.

For the detailed and general feedback, the interviewee's performance in each features is compared against the good interviewees. Feedback is then provided accordingly. The interviewee's performance against the others is also plotted in a graph and serves as a visual feedback. Further, for the general feedback, a score is also computed as an average from all the features of each domain. The general feedback page is shown in figure 7.

7 Limitations and future work

The main limitation of this work is the lack of a deeper connection of the lexical, prosodic and facial analysis. On the one hand this was done intentionally to have a more explainable and understandable system for the end-user but on the other hand a global analysis would probably yield better results, since it could also learn from the interplay between low-level features from the different dimensions. However, a data set with the labels that would be required for the training of such a system might be difficult to find or create. Further, there is is no rigorous way to asses the generalizability of the models to the MIT dataset. The current models

[3] https://gitlab.lrz.de/lab-courses/xai-lab-ws-2022/tobias/critical-speech-analysis-with-explanations/-/raw/main/screenshot$_a nd_d emo/dashboard_d emo.m4v$

[4] https://flask.palletsprojects.com/en/2.0.x/

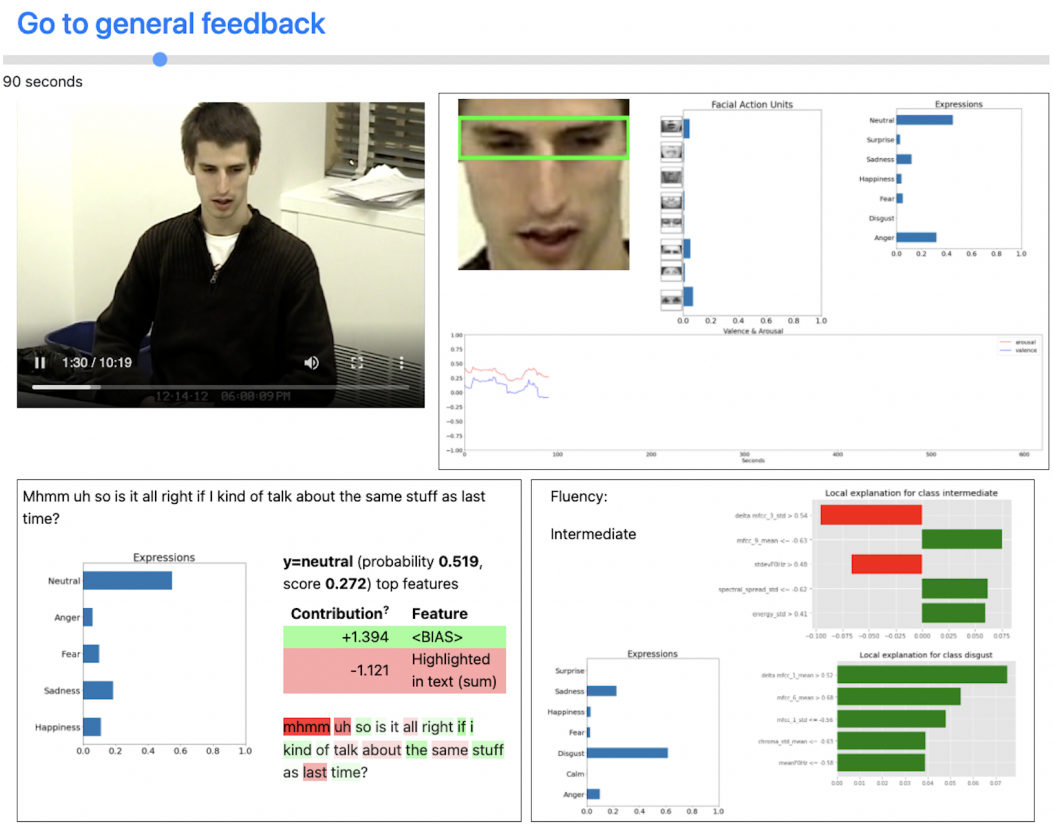

Figure 6: Overview page of the dashboard.

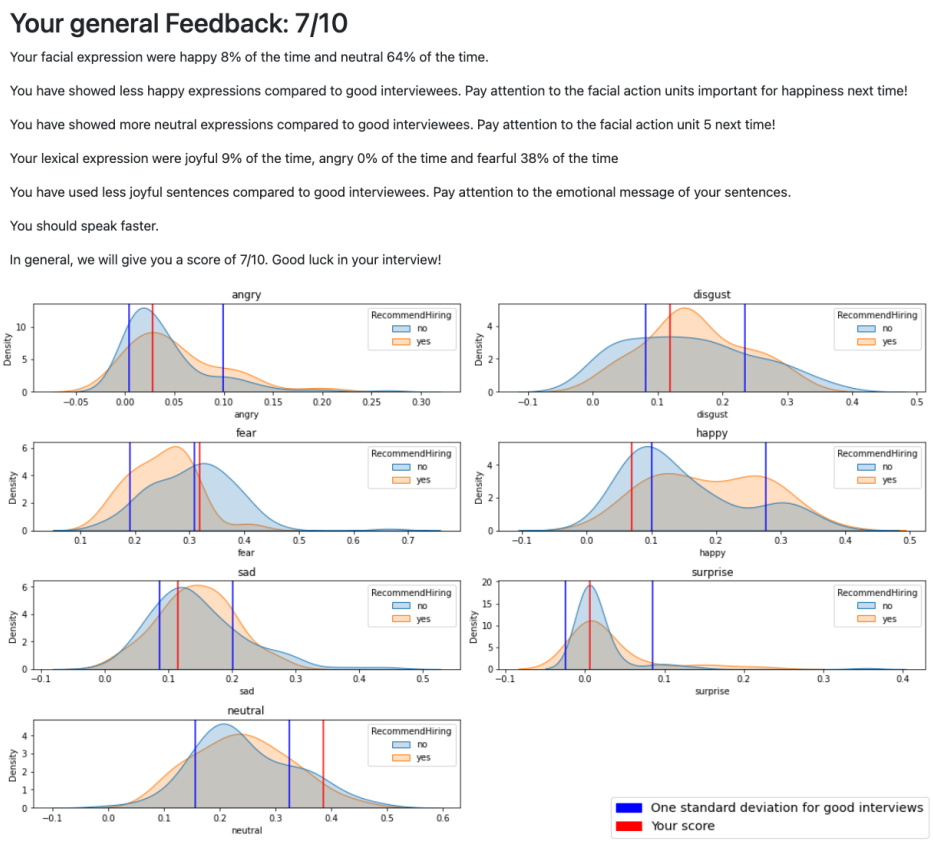

Figure 7: General feedback page of the dashboard.

were trained on non-job-interview data sets where the appropriate labels are given. It is possible that the models perform much worse in an interview context. Although a manual inspection of the interviews with analysis of the feedback and the scores proves this unlikely. So far the system mainly focuses on the emotions and other qualities that are not directly related to how good an interviewee fits professionally into a position. While this is much harder to measure, a first approach could include topic modeling and fact checking, i.e., how much does the interviewee stay on the topic and are the things that he/she says plausible for the interviewer. Questions like these will have to be answered by future systems.

8 Conclusion

In this work, a prototype dashboard is developed for the evaluation of an interviewee's performance and proposed ideas for providing meaningful feedback as a basis for further improvements. The task consists of three different domains - lexical analysis, prosodic analysis and facial analysis - and extracted features from each, where the emotions displayed by the interviewee are the main focus. Furthermore, machine learning models are trained and statistical analysis is applied to investigate the association between these features and the quality of an interview, which enables us to give explanation and feedback to the users. This system can be further extended by additional features, e.g. topic modeling, and will provide support automating the evaluations of job interviews.

References

Speech-to-text documentation nbsp;—nbsp; cloud speech-to-text documentation nbsp;—nbsp; google cloud.

Tadas Baltrusaitis, Amir Zadeh, Yao Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. pages 59–66.

Joan-Isaac Biel, Lucía Teijeiro-Mosquera, and Daniel Gatica-Perez. 2012. Facetube: Predicting personality from facial expressions of emotion in online conversational video. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, page 53–56, New York, NY, USA. Association for Computing Machinery.

Paul Boersma and David Weenink. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.1.38, retrieved 2 January 2021 http://www.praat.org/.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

Didan Deng, Zhaokang Chen, and Bert Shi. 2020. Multitask emotion recognition with incomplete labels. pages 592–599.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Kate Dupuis and M Kathleen Pichora-Fuller. 2010. Toronto emotional speech set (tess)-younger talker_happy.

Paul Ekman and Wallace V. Friesen. 1978. Facial action coding system: a technique for the measurement of facial movement.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *Computational Linguistics and Intelligent Text Processing*, pages 152–165, Cham. Springer International Publishing.

Theodoros Giannakopoulos. 2015. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12).

Allen I Huffcutt, James M Conway, Philip L Roth, and Nancy J Stone. 2001. Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, 86(5):897.

Philip Jackson and SJUoSG Haq. 2014. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Iftekhar Naim, Md. Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2018. Automated analysis and prediction of job interview performance. *IEEE Transactions on Affective Computing*, 9(2):191–204.

James Pennebaker, Roger Booth, Ryan Boyd, and Martha Francis. 2015. Linguistic inquiry and word count: Liwc2015.

Alan Preciado-Grijalva and Ramon F Brena. 2018. Speaker fluency level classification using machine learning techniques. *arXiv preprint arXiv:1808.10556*.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

KR. Scherer and HG Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning.