DeBIT: De-Biasing Interviews with Transcripts

Stanford CS224N Custom Project

Yu-Ann Wang Madan
Department of Computer Science
Stanford University
yuann@stanford.edu

Yuxin Ding
Department of Computer Science
Stanford University
yd2406@stanford.edu

Kate Shijie Xu
Department of Computer Science
Stanford University
skxu@stanford.edu

Abstract

Conventional wisdom advises interviewees to speak clearly and utilize their body language to make a good impression on the interviewer. However, these metrics are subjective and may be influenced by bias. There have been several studies on how to de-bias interviews including offering the option of phone vs. online calls and standardizing interviews. For this project, we aim to analyze audio recordings of interviews to detect (and then potentially remove) biases associated with auditory and visual cues.

We trained a model with wav2vec2 embeddings on the MIT interview dataset [1] to predict interviewee ratings. Since our focus is on isolating visual cues and audio cues which might lead to bias, we engaged with MTurk to relabel and re-evaluate the interviewees, and trained a model to predict interview performance based on audio only. We then manipulated the pitch and accents of these audio files and re-evaluated them with our model to see if our model was biased against genders and accents. Ultimately there wasn't enough bias in the labeled data across gender and accents, and our model had issues predicting different values with the embeddings. We think this opens up the opportunity to explore other types of ratings (e.g., recommend to hire versus overall rating), other biases, and potentially other model architectures in the future.

1 Key Information to include

- Mentor: Ethan A. Chi
- External Collaborators (if you have any): None
- Sharing project: N/A

2 Introduction

From the minute you submit your resume, bias exists in the recruiting process. Interviewees are subjected to the conscious and unconscious biases of the interviewer, and whether someone gets hired may very well depend on the interviewer rather than on the candidate's core substance. In recent years, more and more companies are recognizing the need to debias the recruiting process to hire the best employees. For example, interviewers at Google are trained to minimize their unconscious biases and ask structured questions to evaluate candidates on a set of predetermined rubrics. The hiring decision is also determined by a panel of hiring committee rather than by the interviewer.

However, just as radiologists could benefit from machine learning models taking an initial screen (or second look) at x-rays for cancer, machine learning models could be beneficial for offering a second opinion in behavioral interviews. In the industry, companies such as Cangrade and Headstart uses AI to identify pools of good candidates, but there is yet to be a product that uses ML to automate the hiring decision process. Meanwhile, in the academic world, there's been some work to demonstrate what matters during an interview (Naim [1] showed excitement, engagement, and friendliness were the most important attributes to getting a highly scored interview), there's also been some work for detecting bias in written work (bias in academic peer reviews in ICLR submissions[2], gender bias in recommendation letters for professors[3]) as well as bias in audio (police prosody[4]). However, there's been sparse ML research on creating a model to predict interview outcomes, and then training the model to minimize potential biases. This task is what we hope to explore in this project.

3 Related Work

In Naim [1], the researchers took 138 MIT junior job interviews, and extracted features in facial expression, prosody, and verbal content to see if they correlated with the MTurk workers' overall ratings. The team used an interesting breadth of tools for feature extraction - PRAAT for pitch, vocal intensity, and spectral energy, Shore network for detecting smiles, a constrained Local Model for detecting other facial expressions, and LIWC and LDA for detecting sentiment and content knowledge. They built a model to predict interview ratings, and concluded that "excited", "engagement", and "friendly" were the traits most correlated with the "recommend to hire" prediction, and that prosodic features were the best predictor for these traits.

Naim's's paper inspired us to investigate more on whether there are certain verbal or nonverbal signals we could extract and use to predict the outcome of an interview. There was a MIT study [5] which quantified nonlinguistic communication such as fraction of speaking time, engagement measure (turn taking dynamics modeled using a hidden Markov model (HMM) and measuring the coupling between two of these HMM's to estimate the influence people have on each other), emphasis (using the standard deviation of formant frequency, spectral energy, and energy in frame), and mirroring. This study concluded that speaking time and emphasis were better predictors of overall ratings. Lei Chen from Rakuten [6] also did an interesting study on structured monologue interview questions and created a new framework (BARS) with industrial psychologists for humans and machines to predict interview performance. Chen used Emotient FACET SDK to detect and extract facial expressions, used doc2vec to provide content related measurement, and used LIWC for lexical features and applied all of these to an SVM regression.

A comment that stood out from Naim's paper was that women were penalized more for pauses and filler words than men, and this inspired us to look more into biases in interviews or other forms of human evaluations. Manzoor [2] showed that there is affiliation bias in ICLR peer reviews by using a Naive Bayes classifier and calculating the AUC between a double blind year versus single blind year for the same paper. Madera [3] showed that faculty recommendation letters for women were more likely to reflect "doubt raising" words and content. Camp [4] also used prosody to determine that police officers spoke differently to white versus black motorists using human raters (university students as well as DMV participants).

The aforementioned related works helped us shape the goals of our project - we want to create a machine learning model which could use audio recordings to predict interview ratings, and identify and eliminate certain forms of bias.

4 Approach

For our baseline, we created a linear regression model (code here) to see if we could predict the audio interview ratings. Our inputs were mel-frequency cepstral coefficients (mfccs) extracted with librosa at a sampling rate of 16000 Hz. We calculated loss with mean square error. Initially to deal with the high losses, we also introduced an Adam optimizer and larger batch size of 128.

For our model, we designed the model based on the encoder-decoder architecture. We used a wav2vec2 feature extractor as the encoder. It's a transformer based feateure extractor that takes in log-mel features for each audio snippet and outputs 256 outputs feature vector. The decoder is designed in-house. We choose fully connected layers as the base architecture since the model predicts a single

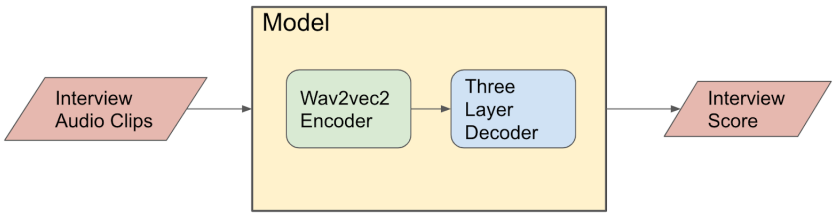


Figure 1: Interview predictor model architecture

interview score. We performed a naive model architecture search to find a suitable architecture. Within the model architecture search, we tried varying the number of layers, the activation functions and hidden dimensions while keeping other parameteres the same.

To detect bias (both gender and accent based), we used the gender swapped audio as our new inputs to see if our model outputs differed from (1) the initial Turkers' ratings on the non-altered files and (2) our model's own outputs on the non-altered files. Using a similar approach as Manzoor's ICLR bias detection [2], our equation for measuring bias (γ) was:

$$\gamma = (\hat{y}_{altered} - \hat{y}_{original}) * 1[(\hat{y}_{altered} - \hat{y}_{original}) > valerror_{original}] \quad (1)$$

Where $\hat{y}_{original}$ is the model predicted rating of the original audio, $\hat{y}_{altered}$ is the model predicted rating of the altered voice, multiplied by an indicator function to compensate for some validation error existing in the model.

5 Experiments

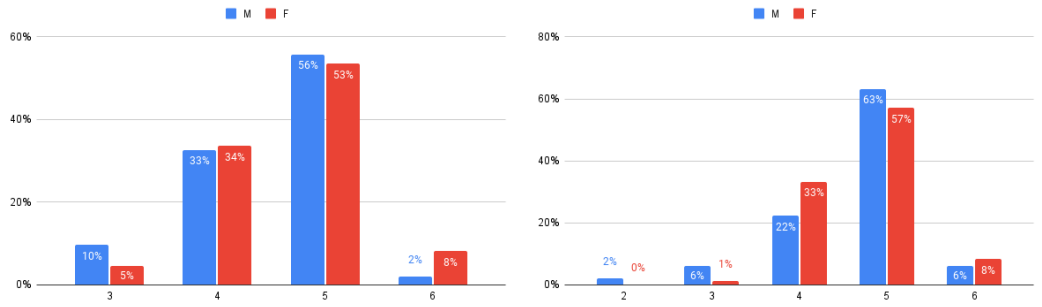
5.1 Data

Data relabeling, processing, and preparation For this study, we used the audio portion of the MIT interview dataset [1]. The interview dataset consists of 138 recordings of juniors from MIT describing their work and school experience, leadership ability, and weaknesses. The original dataset had MTurkers watch each video interview and then rate the candidate across a handful of attributes. Since we're only interested in the audio recordings of the interviews, and we know that smiling, physical appearance, and body language may have influenced the original dataset's labels, we re-evaluated the dataset with a new group of MTurkers. These 5 MTurkers were given audio recordings of the candidates' interviews, and rated their impressions of the candidates on a scale of 1 (least x) to 7 (most x). We chose a subset of labels from the the original article because they were more applicable to audio-transcript data: overall rating, recommend hiring, engaging in tone, excited, friendly, calm, not stressed, authentic, and not awkward.

Initial take on bias After we gathered the new MTurk ratings, we wanted to do a comparison between the video ratings and audio ratings across gender.**[Figure 2]** In our initial comparison between the overall ratings, we noticed the general shape of the distribution was similar, but audio ratings were slightly skewed with lower ratings. Splitting the analysis by gender, we saw the distribution of ratings between males and females were largely similar in video (in fact, 8% of females received a score in the 6 range versus only 2% of men). Applying the same analysis to our audio only ratings, we noticed shifting to audio slightly worked in the men's favor: their distribution of lower scores (2, 3, 4) moved to 30% versus previous 43%, and their share of higher scores (5,6) shifted to 69% from 58%. The women's share of lower scores also decreased (from 39% to 34%), but their high scores did not experience as high a jump (65% of women received the high scores versus 69% of men).

Augmenting audio data In order to identify possible sources of bias, we augmented the existing dataset by swapping the gender and by changing the accents of the audio recordings.

For gender swapping, we first manually labeled the gender of each candidate, then used librosa to alter the pitch of the audio recordings so that the candidates sounded as if they're of the opposite gender. We noted that in each interview, the interviewer and the interviewee are of the same gender, so altering the pitch of the entire recording will change the gender of both parties.



(a) Video interview ratings split by gender

(b) Audio interview ratings split by gender

Figure 2: Showing the change in score distribution between males and females, video versus audio interview ratings

For accent alteration, we used gtts (Google Text-to-Speech) to read the transcripts of the audio recordings in different English accents. We chose Australian, British, Canadian, and Indian accents because they sounded distinct from each other.

Preprocessing: Chopping up the Audio Since the audio clips can range from 5-10 minutes, we decided to chop the clips into 30 seconds chunks that have 10 second overlaps. This made the datasets easier to load and process.

5.2 Evaluation method

Model evaluation: The model is trained with L2 loss and is evaluated also on such loss. The model quality is evaluated based on its loss on the validation and test set.

Bias evaluation: The bias of the model is evaluated based on our model bias evaluation equation we proposed.

5.3 Experimental details

For the model, we used the hugging face library with the following hyper-parameters:

- batch size = 64
- learning rate = 3e-5
- warmup ratio = 0.1
- epochs = 50
- activation function: ReLU
- optimizer: AdamW
- warm start checkpoint: wav2vec2 model trained on 960 hours of Librispeech.

5.4 Results

5.4.1 Baseline

For our baseline model, after 8000 epochs our training loss was at 0.1599 and our validation loss was at 0.51160. We then tried to predict the ratings of the swapped audio files to see how they compared with the actual ratings. Unfortunately, our model largely predicted close to the same value ($\sim$5.07) **[Figure 3]**. The model prediction for the same set of data (unaltered) also landed around 5.07. This showed us that a linear regression with MFCC was not effective at predicting the nuances in ratings between different candidates. This also suggested a potential issue with our labeled data.

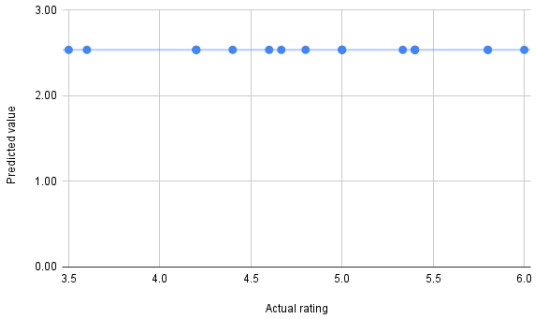


Figure 3: Comparing the predicted results of gender altered audio versus the actual ratings of the original voice

5.4.2 Decoder Model Architecture Search

We decided to use the wav2vec2 feature extractor as the model encoder, but had to design the model decoder by ourselves. As the most straightforward solution, we started with fully connected models and performed model architecture search. We experimented by varying the number of layers, activation function and hidden dimensions in the decoder while keeping other model training parameters the same. The experiment results are shown in Table. 3.

Num of Layers	Activation Function	Hidden Dimensions	Validation Loss	Test Loss
1	-	-	0.498	0.333
3	**ReLU**	**(128, 32)**	**0.484**	**0.295**
3	Sigmoid	(128, 32)	20.610	24.005
3	Tanh	(128, 32)	19.082	22.345
3	ReLU	(256, 64)	0.526	0.307
5	ReLU	(128, 64, 32, 32)	0.496	0.390
5	ReLU	(256, 128, 64, 32)	0.545	0.532

Table 1: Decoder Model Architecture Search Results

Based on the experiment results, the 3 layer ReLU activation model has the lowest loss and, thus, continued the following experiments with such model.

5.4.3 Bias Analysis

We ran inference using our model to predict the ratings of the altered recordings (both gender and accents), and results can be found in **[Figure 4]** and **[Figure 5]**. We saw that there was very small variation among the predictions, and that they were all very close to $\tilde{4}.99$.

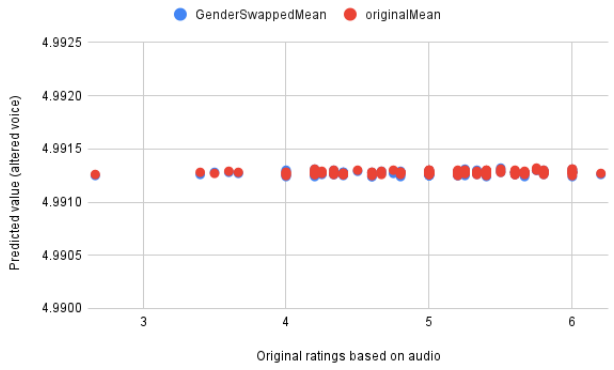


Figure 4: Results on gender swapped audio

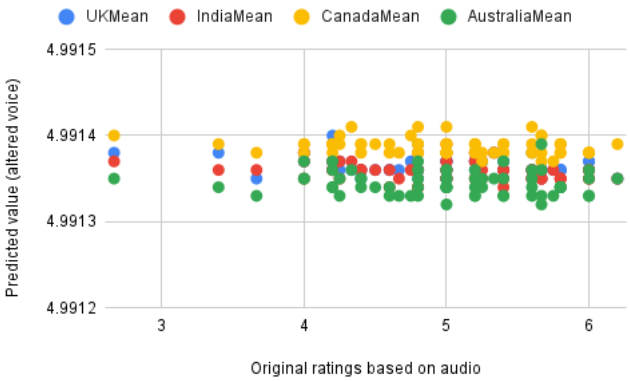


Figure 5: Results on accent altered audio

6 Analysis

6.0.1 Baseline and Decoder Model Architecture Search

It wasn't particularly surprising to us that a Linear Regression model would not be able to predict the interview ratings as there were 128 features used to predict a limited range of ratings (most Turkers rated 3-7). In hindsight, we could have applied dropout and additional layers if this wasn't our baseline model. What was more concerning was a similar thing happened in our Decoder Model. Our hypothesis is that this has less to do with model architecture (although we could try a CNN model with a mel spectrogram as input), but with how the rating data was labeled and how we account and measure for bias in our datasets.
In the future, we could experiment with a binary hire or no hire rating to see if that polarizes the data. We could also increase the breadth of ratings from 1-10, although we suspect most raters will just gravitate towards 5, 6,7,8 based on the behavior we witnessed in this round of MTurk ratings. To annotate for bias, we need to be more precise on which traits have a statistically significant difference in ratings. We used gender and accents because it was easily detected on audio, but the rating distribution was very similar (and the dataset heavily skews toward North American accented English). Finally, we analyzed gender at a very high level, but we could have linked it to other NLP attributes in the interview (e.g., profanity, fillers, content knowledge and action based words), and see if one gender is rewarded/penalized for it more than another. For example, if we used recommend to hire, in a preliminary analysis we looked at the relationship between overall rating vs. recommend to hire [Figure 6]. It was harder to predict a woman's hire score given her overall interview score (R^2 is lower) and as the overall rating gets higher, a male versus female's hire rating diverges (male is higher).

6.0.2 Bias Analysis

As mentioned in the results section, regardless of the altered recordings that were used as input, our model generated predictions that were very close to 4.99. As can be seen in the tables below, the percent difference between the predictions based on the altered recordings and the predictions based on the original recordings were very tiny and insignificant to make any valuable judgment on potential model biases.

Nonetheless, for gender swapped recordings, we see that ratings for recordings that were altered from male to female, scored higher on average than the recordings that were altered from female to male. In fact, the percent difference of the final 'female' recordings was smaller than that of the final 'male' recordings. This can be explained by the fact that female candidates performed better in our audio-based Mturk labels.

For accented readings, we see that the average rating for recordings with Canadian accent was higher than that of the other accents. British accent performed the second best, followed by Indian, then Australian. This can be due to the fact that the original audio recordings were mostly spoken in American/Canadian accent, thus the model might have picked up cues that the Canadian accent was

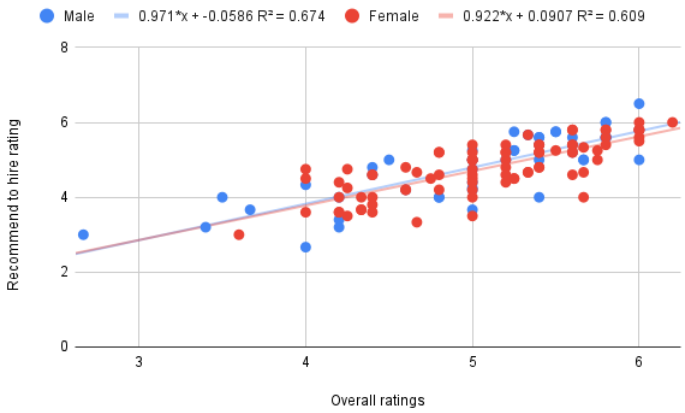


Figure 6: Comparing the correlation of overall scoring with recommend to hire, male vs. female

Altered Gender	Average Prediction	% difference
Female	4.99127	-0.00011
Male	4.99126	-0.00015
Overall	4.99127	-0.00010

Table 2: Percent difference by accents

correlated with a higher score. Again, the percent difference between the predictions based on the original and the altered recordings was too small to say that there is significant accent-based bias in our model.

Accent	Average Prediction	% difference
Australian	4.99134	0.00133
British	4.99136	0.00173
Canadian	4.99139	0.00217
Indian	4.99136	0.00160

Table 3: Percent difference by accents

6.0.3 Raters

Finally, Naim [1] raised this concern in his study as well, but we used MTurkers for our interview ratings. We paid for the premium vetted version of MTurk, however, this population is still different from a job recruiter, who might assess a candidate differently. If we had wanted to invest more time and money, we could have worked with Prolific or Upwork to recruit recruiters for ratings.

7 Conclusion

Overview In this project, we created a model that can use audio recordings of interviews to predict the interview rating. We were hoping that the model would perform differently depending on the speakers' gender or accent in the audio recording which would reveal implicit bias in the model itself, but the model ratings stayed consistent regardless of variations in the audio recording.

Key takeaways Overall, we still believe it is possible to predict interview ratings using audio and detect bias in the data and the model. However, we needed to be more thoughtful of how we characterize the ratings (and how differentiated they are), and be more careful on how we annotate bias to make sure there actually was bias there. With some revisions to our data labeling practice and more care selection of bias traits, we think we could have uncovered more interesting results.

Reflection and future improvements The main problem that led to the inconclusiveness of our results is that what we're trying to predict (hire rating) has a small range from 1-7, and this caused our model to predict the same (or nearly the same) rating each time per candidate. Given more time, we would look more into other ratings (e.g., recommend to hire) or traits (e.g. "engaging", "excited", "authentic") to see if any has a wider spread across candidates, which would hopefully give more variation to our model output and give us more to significant differences to analyze.

Another possible improvement we can do is to experiment with other model architectures, such as CNNs for logmel spectrograms or more complex decoders. Our current model architecture may not be effective against predicting interviewee result, but we were constrained by computational and time resources, and experimented with as many configurations as we could.

Finally, we were only able to test out potential gender based or accent based biases. Given more time, we could use more forms of audio manipulation to identify potential biases. For example, the Google Text-to-Speech API offers more than 200+ voices across 40+ languages and variants, and this can help with identifying possible voice or accent based biases. We can also use time-stretch to manipulate the speed of an audio recording and potentially identify age and speaking speed based biases. We can even adjust the tone of an audio recording to simulate different emotions, and quantify the influence of emotion.

References

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

[2] Emaad A. Manzoor and Nihar B. Shah. Uncovering latent biases in text: Method and application to peer review. *CoRR*, abs/2010.15300, 2020.

[3] Juan M. Madera, Michelle R. Hebl, Heather Kimberly Dial, Randi C. Martin, and Virgina Valian. Raising doubt in letters of recommendation for academia: Gender differences and their impact. *Journal of Business and Psychology*, 34:287–303, 2019.

[4] Nicholas P. Camp, Rob Voigt, Dan Jurafsky, and Jennifer L. Eberhardt. The thin blue waveform: Racial disparities in officer prosody undermine institutional trust in the police. *Journal of Personality and Social Psychology*, 121(6):1157–1171, 2021.

[5] Vikrant Soman and Anmol Madan. Social signaling: Predicting the outcome of job interviews from vocal tone and prosody. *IEEE In'tl Conference on Acoustics, Speech and Signal Processing*, 2009.

[6] 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. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 161–168, New York, NY, USA, 2016. Association for Computing Machinery.

A Appendix (optional)

Code links:

1. MTurk audio interview ratings: csv
2. Data augmentation/accent generation code: code
3. Data pre-processing code: code
4. Baseline Linear Regression with MFCC: code
5. Model training code: code
6. Model evaluation code: code
7. Model inference on altered audio: code