# Fairness in Rating Prediction by Awareness of Verbal and Gesture Quality of Public Speeches

Rupam Acharyya *,[1] Ankani Chattoraj *,[1] Shouman Das [1] Md. Iftekhar Tanveer [2] Ehsan Hoque [1]

University of Rochester[1], Spotify Research [2]
racharyy@cs.rochester.edu, achattor@ur.rochester.edu, sdas13@ur.rochester.edu,
go2chayan@gmail.com, mehoque@gmail.com

**Abstract**

The role of verbal and non-verbal cues towards great public speaking has been a topic of exploration for many decades. We identify a commonality across present theories, the element of "variety or heterogeneity" in channels or modes of communication (e.g. resorting to stories, scientific facts, emotional connections, facial expressions etc.) which is essential for effectively communicating information. We use this observation to formalize a novel HEterogeneity Metric, HEM, that quantifies the quality of a talk both in the verbal and non-verbal domain (transcript and facial gestures). We use TED talks as an input repository of public speeches because it consists of speakers from a diverse community besides having a wide outreach. We show that there is an interesting relationship between HEM and the ratings of TED talks given to speakers by viewers. It emphasizes that HEM inherently and successfully represents the quality of a talk based on "variety or heterogeneity". Further, we also discover that HEM successfully captures the prevalent bias in ratings with respect to race and gender, that we call sensitive attributes (because prediction based on these might result in unfair outcome). We incorporate the HEM metric into the loss function of a neural network with the goal to reduce unfairness in rating predictions with respect to race and gender. Our results show that the modified loss function improves fairness in prediction without considerably affecting prediction accuracy of the neural network. Our work ties together a novel metric for public speeches in both verbal and non-verbal domain with the computational power of a neural network to design a fair prediction system for speakers.

## 1 Introduction

A great talk depends on how efficiently the inner thoughts of the speakers are expressed to an audience [14, 39]. A good talk requires encompassing ethos, pathos and logos [43] as well as including variety, such as, adding humor, asking questions, using quotations, drawing analogies etc [45, 19, 20, 8, 52, 15]. For example, celebrity chef Jamie Oliver gave an award winning TED speech in 2010 where he attempted to convince a diverse audience to change their most basic eating habits. After establishing credibility, he showed compelling statistics confirming that death from diet related disease is more prevalent than any other diseases, accidents or murder. He ends with easy alternative solutions to encourage healthy eating [41]. Different aspects of the talk resonated with different people, depending on their age, life style, experiences and cultural background, making it an awarding winning speech. In addition to verbal, nonverbal components of a talk also play a key role in determining its appeal [29]. Effective use of both transcript and gesture and a deliberate alteration of message through verbal and nonverbal cues give shape to the main message of a speech [1, 51].

One common thread that ties the important aspects of a good public speech in both verbal and non-verbal domain is "variety or heterogeneity": the efficient way of conveying the main message by providing information in the form of stories, scientific facts, emotional connections, personal experience etc. We therefore formalize the idea of variety and heterogeneity in both verbal and non-verbal domain of a public speech by defining a novel, "HEtero-geneity Metric" ($HEM$). We conduct our investigation on a diverse set of talks and ratings as found in the TED talk website. The speakers and the viewers are from different cultures, age groups, and backgrounds. We show that $HEM$ has a meaningful relationship with the ratings of TED talks in both the verbal and non verbal domain. This emphasizes that

[*] equal contribution

$HEM$ indeed quantifies the quality of a talk based on heterogeneity/variety. We find that the desirable (positive rating, e.g., fascinating) and undesirable (negative rating, e.g., unconvincing) ratings of talks grow and shrink respectively with the increase in $HEM$. This holds for both the verbal and non verbal domain. However, as expected, we also observe that too much variability (higher values of $HEM$) can be overwhelming and distracting for the audience causing a decrease (increase) in positive (negative) ratings.

Interestingly, we also find that our novel metric, $HEM$, captures discrepancies in rating of TED talks w.r.t race and gender. Motivated by this observation, we introduce *fairness by quality* in designing a public speech rating prediction model. Our prediction model is a neural network with a modified loss function which aims to reduce discrepancy in ratings for talks with comparable quality as quantified by $HEM$ besides improving prediction accuracy. In summary: (1) we formalize a novel metric $HEM$, to quantify quality of a talk in both verbal and non-verbal domain based on variety or heterogeneity. (2) we show a meaningful relationship between $HEM$ and rating labels (both positive and negative) of TED talks which justifies its credibility as a quality measuring metric. (3) we show that $HEM$ successfully captures the inherent bias in rating with respect to sensitive attributes, gender and race (4) finally we incorporate the $HEM$ into the loss function of a neural network to design a fair rating predictor for public speeches w.r.t. *race* and *gender*, shown in Figure 1. To the best of our knowledge, this is the first work that uses a "multi-modal" novel metric to build a fair prediction model.

## 2 Related work

Public speaking has been used as a tool to reach out to masses for centuries. A good talk requires credibility or trustworthiness, emotional appeal and finally logic or reasoning [6, 43]. In recent years, with a more multi-cultural and age diverse audience [16, 52, 15], it has been proposed that adding humor, asking a question, using analogies and quotations can be effective ways of connecting to the audience [38, 54, 30, 12, 46]. In fact one important point of speech editing demands addition of "variety" to increase its appeal [12, 10].

Research shows that majority of all human communication is non-verbal [37, 32, 33] and basic facial expressions of emotion are similar across cultures [18]. For example we typically smile while speaking about happy events, we open our eyes wide with astonishment, we raise an eyebrow in suspicion or open our mouth to express surprise, we pucker our lips to express disapproval, disappointment or concern. Because of such connections between emotions and facial expression, they form an integral part of delivering a good speech [35, 21, 22, 50, 4].

All these ideas are stitched together by the fact that a good speech requires "variety and heterogeneity" both in the verbal and non-verbal domain. However, too much heterogeneity in transcript or gestures can be detrimental because if there is too much to think about, it can cause cognitive overload as we try to cope ("bounded rationality") [49]. Similarly, too much variability in facial expressions can be interpreted as nervous gestures (Chapter 12 of [47]).

Mitigation of bias in AI models has drawn much interest in recent times as automated systems are now being used in sensitive decision making such as criminal justice systems [53], filtering resumes for job applicants [40, 13, 36] etc. There have been substantial amount of research in recent times on the notion of "fairness" and its improvement in machine learning models [11, 17, 23, 48, 24, 25]. A recent survey about fairness in machine learning is very well captured in [32]. However, there has not been much work of incorporating fairness in the domain of public speech rating prediction. In a recent work [2] the authors used counterfactual fairness [27] for predicting the rating of public speaking in TED talks. Their work promotes similar ratings for individuals with comparable transcript and skill set in a counterfactual world without any mention of explicit quality quantifying metric for public speech. Moreover, their work focuses on fair predictions in the verbal domain only. We however, take a complementary approach of first introducing a psychology motivated metric ($HEM$) to quantify the quality of a talk and then using it to identify the existence of bias in TED talk ratings. We further incorporated it into our model for building a fair prediction system. Our work is the first of its kind in combining psychology and computationally efficient models with multi-modal information to build a fair prediction application.

## 3 Data

**Data Acquisition**: We collected the public speaking data from TEDtalk website [1]. The dataset contains speeches published between 2006 and 2017 from 1980 talks (after removal of outliers) covering $>$ 400 categories such as

[1] Videos on `www.ted.com`

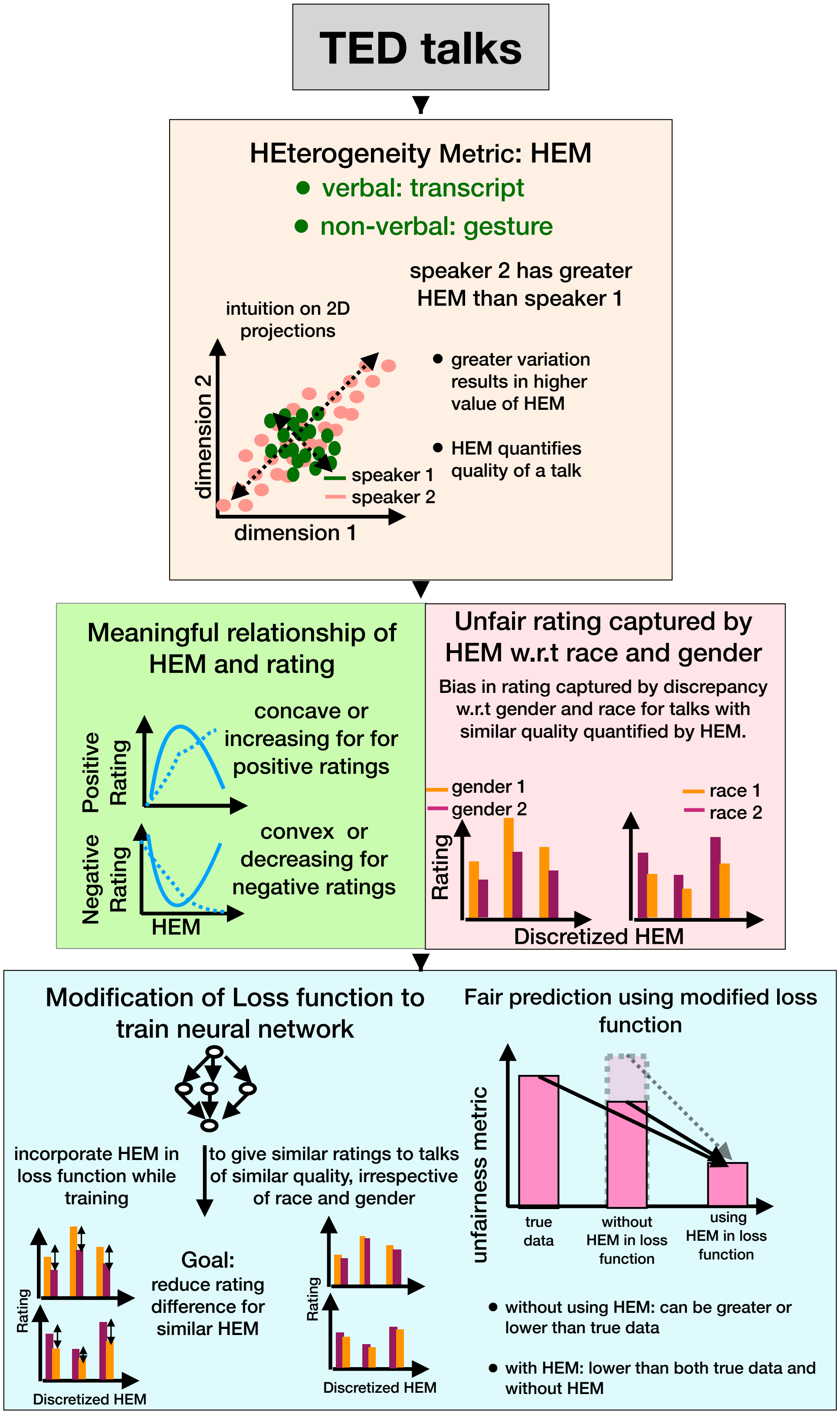
TED talks
HEterogeneity Metric: HEM
verbal: transcript
non-verbal: gesture
intuition on 2D projections
speaker 2 has greater HEM than speaker 1
greater variation results in higher value of HEM
HEM quantifies quality of a talk
dimension 2
dimension 1
speaker 1
speaker 2
Meaningful relationship of HEM and rating
Positive Rating
concave or increasing for for positive ratings
Negative Rating
convex or decreasing for negative ratings
HEM
Unfair rating captured by HEM w.r.t race and gender
Bias in rating captured by discrepancy w.r.t gender and race for talks with similar quality quantified by HEM.
gender 1
gender 2
race 1
race 2
Rating
Discretized HEM
Modification of Loss function to train neural network
incorporate HEM in loss function while training
to give similar ratings to talks of similar quality, irrespective of race and gender
Rating
Discretized HEM
Goal: reduce rating difference for similar HEM
Rating
Discretized HEM
Fair prediction using modified loss function
unfairness metric
true data
without HEM in loss function
using HEM in loss function
without using HEM: can be greater or lower than true data
with HEM: lower than both true data and without HEM

Figure 1: **Fair prediction pipeline using $HEM$ metric**

Table 1: Breakdown of TED talk Dataset w.r.t. gender and race

| Property | Sub Property | Quantity |
|---|---|---|
| Total Number of talks | | 1980 |
| Gender of Speaker | Male | 1307 |
| | Female | 659 |
| | Other gender | 14 |
| Race of Speaker | White | 1573 |
| | African American | 147 |
| | Asian | 173 |
| | Other Race | 87 |

Table 2: Pre-processed TED Dataset Attributes

| **Sensitive attributes** $S$ | $R$: race and $G$: gender |
|---|---|
| **Data attributes** $X$ | $Tr$: transcript $V$: visual gestures and $C$: view count |
| **Label** $Y$ | ratings (3 positive and 3 negative) and $Y$: normalized ratings, $Y^{bin}$: binarized ratings |

science, technology, global issues, health, business entertainment etc. These talks have millions of viewers [2] around the world who come from various background, age and culture. These viewers rate the talks after viewing them based on multiple labels of which we consider 3 positive/desirable rating labels: *fascinating, ingenious* and *jaw-dropping* and 3 negative/undesirable rating labels: *long-winded, unconvincing* and *ok.* We also collected data on the sensitive attributes $S$ (*race* and *gender*) using Amazon mechanical turk following [2]. We use information about the data attributes $X$: *transcript* ($Tr$)*, visual gestures* ($V$)*, view counts* ($C$)*, sensitive attributes* $S$*: race* ($R$)*, gender* ($G$) *and rating label* ($Y$) of each talk to design our prediction model. Summary of the dataset is given in Table 1 (see Appendix). **Extracted Features**: We use transcript (verbal) and visual gestures (non-verbal) for construction of the heterogeneity based metric $HEM$, that is used in all our analyses.

**-Verbal**: We obtain the doc2vec [28] representation of each transcript ($Tr \in \mathbb{R}^d$) by using the `gensim` library [44] and use $d = 200$ for all reported results. We use this to compute $HEM_{tr}$.

**-Non-verbal**: We use OpenFace [5] to extract 17 facial action units. These features, $V \in \mathbb{R}^d$, where $d = 17$, are based on the Facial Action Coding System (FACS) [18] which is widely used in the field of human behavioral analyses. The features are extracted at the rate of 30 frames/sec. We use this to compute $HEM_{ges}$.

**Data Preprocessing**: We normalize the original view counts of the TED talks ($C \in \mathbb{Z}$) using the min-max technique to obtain $C \in \mathbb{R}$ such that they lie in a range of values that comparable to other attributes used in our analyses.

Normalized rating labels $Y$ for each talk was obtained by dividing with its total number of rating counts. It also indirectly neutralizes the effect of how long a video has been online, which we assume gets captured by the total views.

Binary rating labels were computed by thresholding with respect to its median value across all talks, $Y^{bin} \in \{0, 1\}$. The attributes of our dataset are shown in Table 2 in Appendix.

# 4 HEterogeneity Metric (HEM)

Various theories claim that a good talk should have credibility, emotional connection to the audience, logical argument, make use of stories, scientific facts, quotations, humor and so on (see Related work) which we call *characteristics*. Each characteristic when represented by words specific to them would therefore vary with one another. We quantify the verbal quality of a talk by formalizing the heterogeneity across *characteristics* prevalent in the transcript ($HEM_{tr}$).

[2] More than 12 million subscribers on YouTube `https://www.youtube.com/user/TEDtalksDirector`

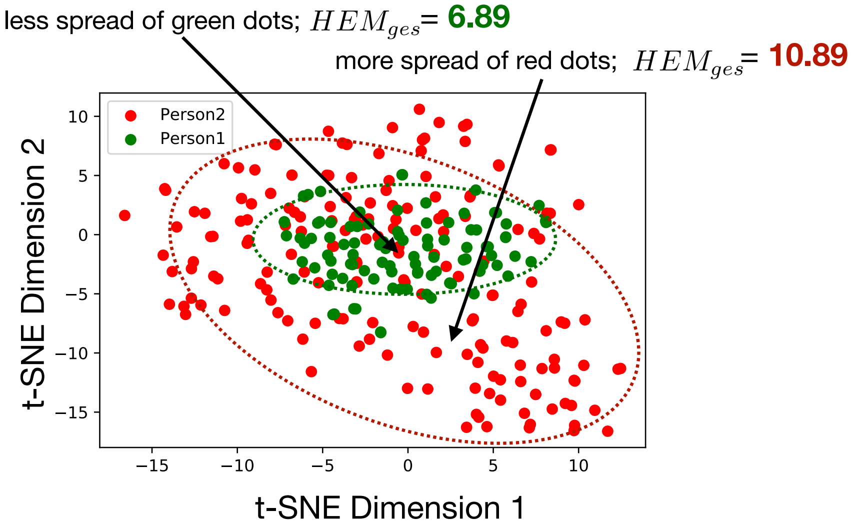

Figure 2: **Gesture pattern of two random speakers**
Speaker represented with red dots has greater variation and hence greater values of $HEM_{ges}$.

Further, multiple studies in the domain of gesture analysis show that the lack of variation in gesture and mannerism makes a talk boring. We quantify the non-verbal quality of a talk by the heterogeneity in facial gestures of a speaker ($HEM_{ges}$). It is important to remember that "quality" here refers to the "impact and effectiveness" of a speech due to heterogeneity as supported by psychological theories (see Related work).

**-Verbal ($HEM_{tr}$):** For each talk we obtained $K$ topics using Latent Dirichlet Allocation (LDA) [9]. Each of these topics is a probability distribution over the set of words in the transcript. We define the summary representation of each topic as the weighted combination (weights are the respective probability) of the Glove embeddings [42] for the highest probable words in that topic. We identify these topics as *characteristics* of the talk by using topic modeling within each transcript. The summary representation of $i^{th}$ topic as $\boldsymbol{T_i} (\in \mathbb{R}^{300})$ is defined by, $\boldsymbol{T_i} = \sum_{j=1}^{n} weight(w_j) \cdot glove(w_j)$, where $n$ is the number of highest probable words chosen to represent the topic and $weight(w_j)$ is the weight of the $j^{th}$ word in the $i^{th}$ topic. We then define the representation matrix of all topics as $\boldsymbol{T} (\in \mathbb{R}^{K \times 300})$ where the $i^{th}$ row is $\boldsymbol{T_i}$. We obtain the topic similarity matrix, $\boldsymbol{U} = \boldsymbol{T}\boldsymbol{T}^{\top} (\in \mathbb{R}^{K \times K})$ which captures the similarity between topics. Heterogeneity metric ($HEM_{tr}$) of a talk is then defined as the product of highest $k$ eigenvalues of $\boldsymbol{U}$. This product defines the volume spanned by the most diverse $k$ topics in the topic vector space (see Theorem 5.2 of [26]). Note that, the more diverse the topic vectors, larger will be the magnitude of the product of top $k$ eigenvalues. This would result into larger volume spanned by corresponding topic vectors. Hence more variety and heterogeneity in a talk with respect to the transcript implies greater value of $HEM_{tr}$. $HEM_{tr} = \prod_{i=1}^{k} \lambda_i$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K \geq 0$ are the eigenvalues of $\boldsymbol{U}$ (all of $\lambda_i$'s are real and non-negative as $U$ is a PSD matrix). In this work, we choose $n = 10, K = 10, k = 5$.
**-Non-verbal ($HEM_{ges}$):**We model the facial gestures using similar technique as in [3] which encodes personalized speaking pattern. OpenFace [5] toolkit is used to extract 17 facial action units related intensity scores from the video at 30 frames/sec. These intensity scores represent various facial muscle movements like inner brow raiser, outer eyebrow raiser, eyebrow lowerer, upper lid raiser, cheek raiser, jaw drop, eye blink etc. Each talk video is divided into 10 second segments with 5 seconds sliding window. For each of these segments, we calculate Pearson correlation coefficient to quantify the similarity among these intensity scores over time. We calculate a total of $\binom{17}{2}$ = 136 correlation coefficients that encodes facial gesture pattern for each 10 second segments. Facial gesture pattern of two speakers are shown in Figure 2. The two colors show gesture pattern from two randomly chosen speakers. Each colored dot is a 2-D t-SNE [31] representation of 136 correlation coefficients that encodes facial gesture pattern for a 10 second segment. The spread of dots of each color shows the amount of heterogeneity present in each speaker's facial gesture. Therefore we define the heterogeneity metric for a speaker's gesture pattern ($HEM_{ges}$) as the maximum distance between two points in 136-dimensional space. Intuitively, this maximum distance gives us the measure of highest variability present in facial muscle movements. For example, person 2 denoted by red dots shows a greater spread

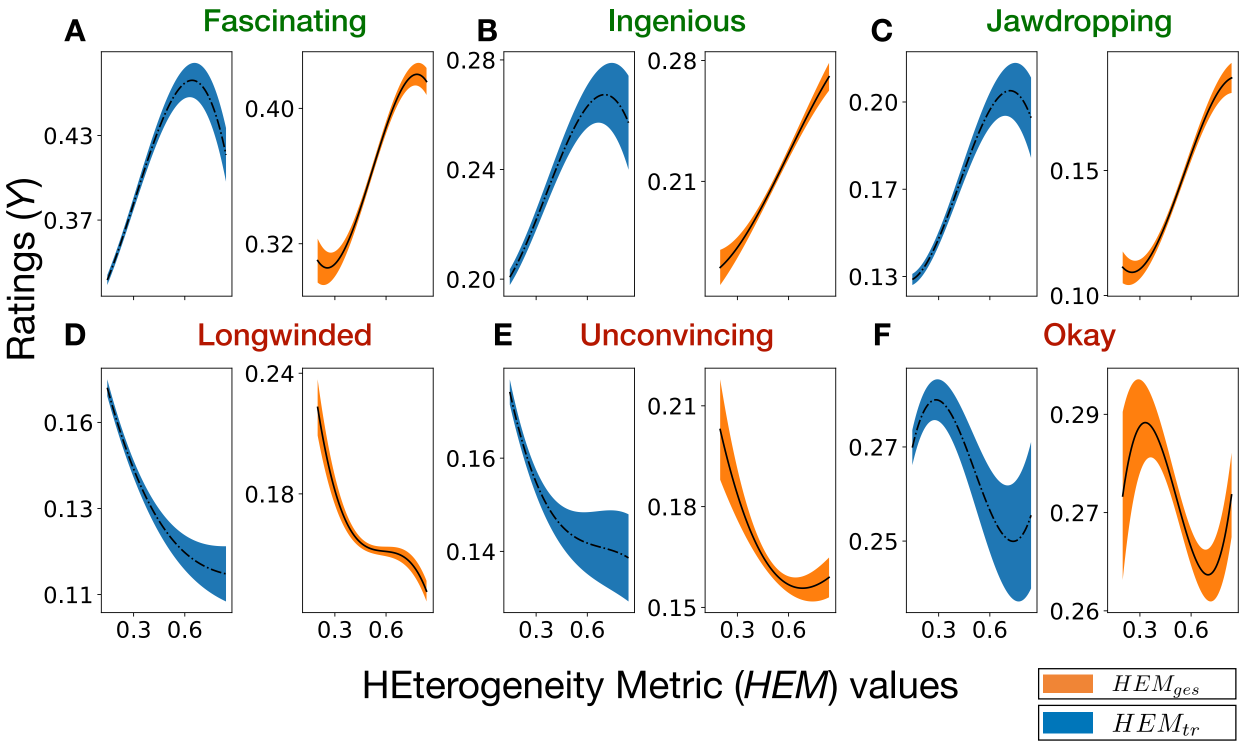

Figure 3: $HEM$ **metric shows meaningful relationship with rating**
**A-C**: Positive ratings show a concave or increasing trend with respect to both $HEM_{tr}$ shown in blue and $HEM_{ges}$ shown in orange. With increase in $HEM$ we expect the quality of speech to improve causing positive rating to increase. However, too much heterogeneity can lead to confusion causing positive ratings to saturate or even decrease. **D-F**: The trend is opposite for negative rating labels, showing a convex or decreasing relationship. Refer to Section 4 for details on choice of rating labels and magnitudes of rating label values.

of 2-D t-SNE values as compared to person 1 denoted by green dots. Correspondingly, the $HEM_{ges}$ for person 2 is 10.89 which is greater than that of person 1 ($HEM_{ges}$ = 6.89). Let the number of 10 seconds segments in a TED talk video be $s$. Denote $p_i (\in \mathbb{R}^{136})$ to be the point representing the correlation coefficients for $i^{th}$ segment. Now, $d_{ij}$ denotes the eucledian distance between points $p_i$ and $p_j$. The $HEM_{ges}$ of gesture pattern for the video is defined as, $HEM_{ges} = \max_{i,j \in \{1,\cdots,s\}} d_{ij}$.

**-Choice of rating labels:** We assume that variety and heterogeneity adds "x-factor" to a speech and determines how engaged and appealing the speech is to an audience. Based on that, we chose 3 positive and 3 negative rating labels which cannot be otherwise trivially attributed to common identifiable causes. Note that the use of the term "quality" associated with $HEM$ is in essence to emphasize the "impact" of a speech on the audience. We do not claim $HEM$ (which represents "heterogeneity" in a speech) is the sole or best way to capture "effective influence" of a speech on audience but it definitely is one useful and relevant way of doing so as shown by the intuitive results of Figure 3 and effectively the first of its kind to the best of our knowledge.

**-Relationship between $HEM$ and ratings:** In order to validate that $HEM$ is an useful metric we investigated the relationship between $HEM$ and ratings. We normalized the $HEM$ values using min-max technique such that they lie in $[0, 1]$. These values are then divided into $5$ equal sized bins and their corresponding mean rating is reported in the X-axis of Figure 3. The Y-axis denotes the corresponding rating value.

If $HEM$ indeed quantifies the variety of characteristics in a transcript and facial gestures of a speaker, then we would expect that its effect on positive rating labels (here, *ingenious, fascinating, jaw-dropping*) would either be a concave function or show an increasing trend. Similarly the effect of $HEM$ on negative rating labels (here, *long-winded, okay, unconvincing*) would be convex or show a decreasing trend. For example, a greater value of $HEM_{tr}$ would indicate use of multiple characteristics of a good speech. This should potentially increase the *fascinating* rating and decrease the *unconvincing* rating. However, too much variety can be detrimental and make the talk annoying to viewers thereby

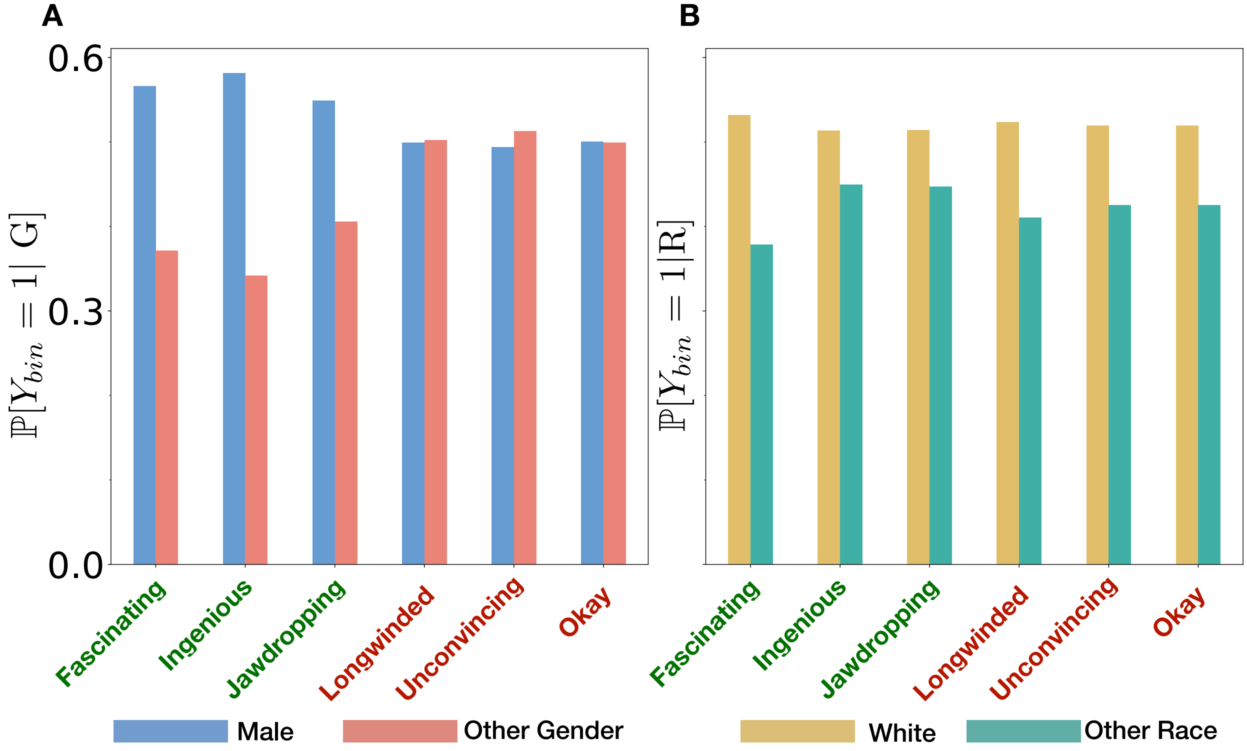

Figure 4: **Rating bias with respect to gender and race**
**A:** Male are more likely to be rated positive labels as compared other gender. No such discrepancy for negative rating labels. **B:** White speakers are more likely to be rated with all 6 labels as compared to speakers of race.

decreasing *fascinating* rating and increasing *unconvincing* rating as shown in Figure 3. Similarly negative rating labels have approximate convex or decreasing relationship with $HEM_{tr}$ and $HEM_{ges}$. This establishes that the psychological intuition motivated novel HEM metric indeed captures an important aspect of speeches that meaningfully influence ratings for both the verbal and non verbal regime. The probability mass of distribution of $HEM_{tr}$ and $HEM_{ges}$ is small for higher and lower values respectively (Figure 11 in Appendix) causing comparably bigger errorbars around those regions. Also note that the normalized rating values are small because of the nature of distribution of ratings across all talks (Figure 8 in Appendix).

## 5 Bias in rating captured by $HEM$

The ratings of TED talks are given by spontaneous visitors to the website who come from different background and are of different age groups. We show that there exists bias in ratings when measured with respect to race and gender. We find discrepancy between $\mathbb{P}[Y_{bin} = 1|\text{ G } = \text{ Male }]$ and $\mathbb{P}[Y_{bin} = 1|\text{ G } = \text{Other gender }]$ for a rating of interest $R$ as show in Figure 4 A. Similar results for race are shown in Figure 4 B. It is important to solely draw attention to the difference in rating probabilities for sensitive attributes in Figure 4 and not individual probability magnitudes. Bias in data can often be counter intuitive as shown in Figure 4 B. It might be trivially expected that white speakers are less likely to be rated with negative labels when compared to speakers of other race. However, we clearly see the data shows opposite trend, highlighting bias here is quantified as any type of discrepancy in rating probability irrespective of assumptions about superior race or gender. We only investigate two cases in Figure 4 to establish the existence of bias in ratings. Several other combinations of sensitive attributes have been explored but not shown here as they only make our claim about prevalence bias in data stronger.

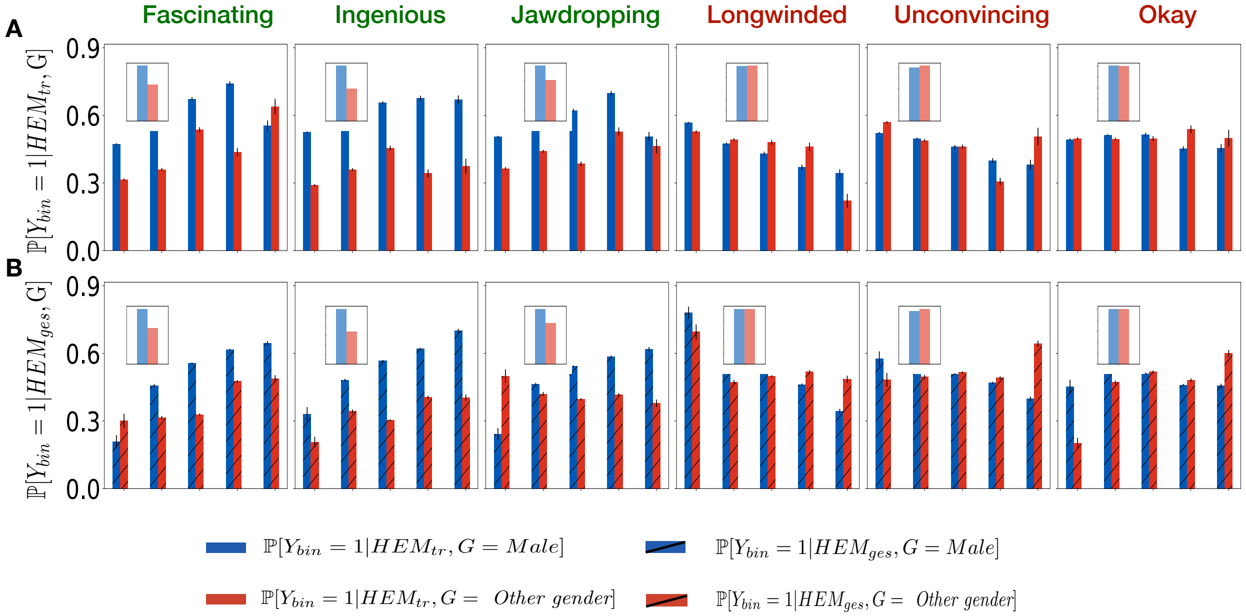

Figure 5: **Significant bias in rating w.r.t gender and** $HEM$
**A:** Significant difference between probability of obtaining a rating for male speakers than for speakers of other gender w.r.t $HEM^{dis}_{tr}$ (trend agrees with true data in Figure 4 A). Negative rating labels do not show significant bias in ratings as observed in true data. **B:** Same as **A** but for $HEM^{dis}_{ges}$. $HEM$ **C:** Significant difference between probability of obtaining a particular rating for white speakers than for speakers of other race w.r.t $HEM^{dis}_{tr}$ (trend agrees with true data in Figure 4 B). **D:** Same as **C** but for $HEM^{dis}_{ges}$.

# 6 Bias in rating captured by $HEM$

We tested whether our novel metric $HEM$ is capable of capturing such discrepancy of TED talk ratings with respect to gender and race. Since $HEM$ captures the "heterogeneity based quality" of the talk, the ratings should be similar for similar quality values, irrespective of race and gender. $HEM \in [0, 1]$ is dicretized into 5 values, $HEM^{dis} \in 0, 1, 2, 3, 4$ by binning $[0, 1]$ into into 5 equal sized bins. We then compute $\mathbb{P}[Y_{bin} = 1|$ G $=$ Male $, HEM^{dis}]$ and $\mathbb{P}[Y_{bin} = 1|$ G $=$ Other gender $, HEM^{dis}]$ for a fixed value of $HEM^{dis}$ as shown in Figure 5 and 6. We find that male and female speakers have significant discrepancy across all positive rating labels for both transcript and gesture shown in Figure 5 and 6. Note that the nature of the bias w.r.t. $HEM^{dis}$ matches existing bias in data. For example, true data shows that male speakers get ingenious rating with greater probability than speakers of other genders (for both transcript and gesture). This trend is nicely captured when computed w.r.t. $HEM$ values. Interestingly, we also observe less discrepancy for negative rating labels which is consistent with observation in true data (for example, ok rating as shown in Figure 4 A). The distribution of $HEM$ values is similar within gender and race ensuring that the metric itself is not biased (Figure 9 in Appendix). It is also important to notice that the bias has a definite pattern across almost all values of $HEM^{dis}$ for a particular rating and sensitive attribute. For example, when comparing w.r.t gender for the rating label jaw dropping, we observe that a male speaker is more likely to obtain that rating across all five values of $HEM^{dis}$. However, we see that for fascinating rating label the trend of bias in rating flips for $HEM^{dis}_{tr} = 4$ as compared to other values of $HEM^{dis}_{tr} = 0, 1, 2, 3$. Similarly, for the rating label jaw-dropping, we find a flip in trend of bias for $HEM^{dis}_{ges} = 0$ when compared to trend $HEM^{dis}_{ges} = 1, 2, 3, 4$. This mismatch in bias pattern for higher $HEM^{dis}_{tr}$ and lower $HEM^{dis}_{ges}$ does not nullify the novelty of $HEM$ but can be explained by negligible probability mass of $HEM^{dis}_{tr}$ and $HEM^{dis}_{ges}$ for values of 4 and 0 respectively (see Figure 11 in Appendix). To test the usefulness and validity of $HEM$ metric, we design a neural network model whose loss function enforces the reduction of rating differences for similar $HEM$ values. Note that the neural network has information about race and gender in the input but the rating discrepancy is minimized only w.r.t. the HEM metric explicitly.

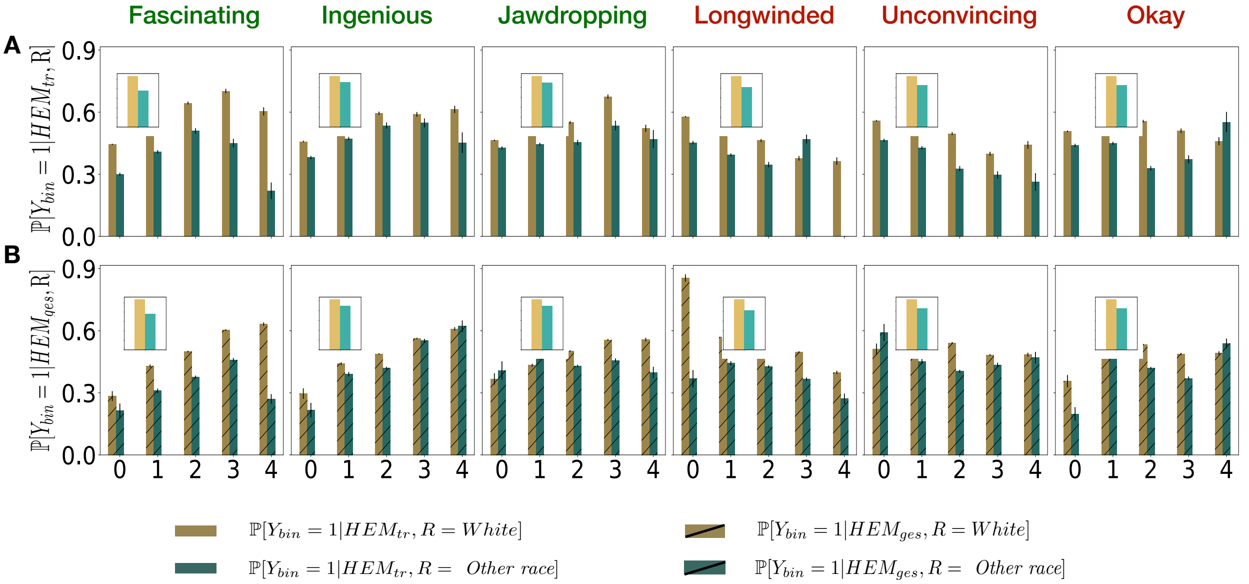

Figure 6: **Significant bias in rating w.r.t race and** $HEM$
**A:** Significant difference between probability of obtaining a rating for male speakers than for speakers of other gender w.r.t $HEM^{dis}_{tr}$ (trend agrees with true data in Figure 4 A). Negative rating labels do not show significant bias in ratings as observed in true data. **B:** Same as **A** but for $HEM^{dis}_{ges}$. $HEM$ **C:** Significant difference between probability of obtaining a particular rating for white speakers than for speakers of other race w.r.t $HEM^{dis}_{tr}$ (trend agrees with true data in Figure 4 B). **D:** Same as **C** but for $HEM^{dis}_{ges}$.

# 7 Fair Model based on multi-modal $HEM$

**-Problem Formulation and Methods:** Each TEDtalk has verbal and non-verbal features (see Section 3). Here the verbal feature is a 200 dimensional vector obtained from the doc2vec representation [34]. To get the non-verbal feature we compute top $k$ (in our case $k = 2$) eigenvalues for the correlation matrix of a segment (See Section 4 for detail about segment) and concatenate these eigenvalues across all the segments. We make this into a fixed length vector by padding 0's at the end. Other than the verbal and non-verbal features we also have the speaker's race and gender information. Input $X^{(i)} (i \in \{1, \cdots, N\})$ is now created by concatenating the verbal feature, non-verbal feature and the sensitive attributes (race and gender) and view counts. We use the binarized ratings $Y^{(i)}_{bin} = \left(y^{(i)}_1, \cdots, y^{(i)}_6\right)$ as the label. Also for each TEDtalk we have calculated its $HEM$ metric as defined in Section 4. Let us denote the overall heterogeneity by $h^{(i)} = (HEM^{(i)}_{tr}, HEM^{(i)}_{ges})$.
Now, our goal is to learn a prediction function $f_\theta$ such that its predicted output is not only close to the ground truth but also has similar prediction for inputs with similar $HEM$ metric value. Formally we need to minimize the following loss function,

$$L(\theta) = Pred(\theta) + \lambda \cdot HEM(\theta) \tag{1}$$

where,

$$Pred(\theta) = \frac{1}{N} \sum_{i=1}^{N} BCE\left(f_\theta(X^{(i)}), Y^{(i)}\right) \ and \ HEM(\theta) = \frac{1}{\binom{N}{2}} \sum_{i,j} ||[|| 2]\hat{Y}^{(i)} - \hat{Y}^{(j)}{}_2^2 \cdot \mathbb{I}\left(|h^{(i)} - h^{(j)}| < \epsilon\right)$$

where $BCE$ denotes the binary cross entropy loss and $\hat{Y}^{(i)} = f_\theta(X^{(i)})$. Here $Pred(\theta)$ represents the prediction loss w.r.t the ground truth and $HEM(\theta)$ controls the discrepancy of the prediction between inputs with similar HEM metric. Here $\epsilon$ and $\lambda$ are the hyperparameters, where $\epsilon$ is the tolerance of $HEM$ difference for which discrepancy in rating is penalized and $\lambda$ controls the strength of $HEM$ loss. Higher values of $\lambda$ forces similar rating for talks with similar $HEM$ value. We use a feed-forward neural network with 1 hidden layer to learn function $f_\theta$ which minimizes

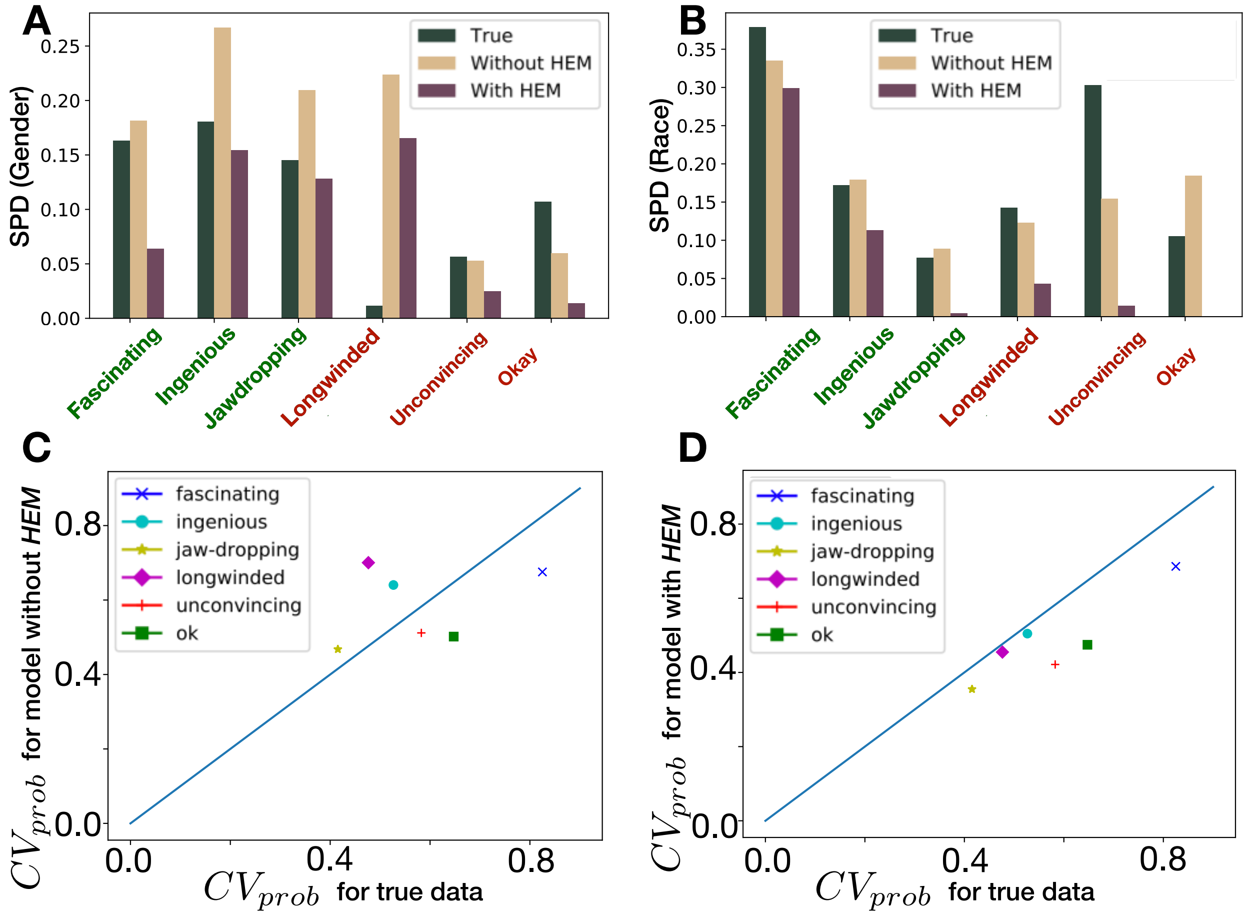

Figure 7: **Significant improvement of fairness by using modified loss function incorporating** $HEM$
**A:** Fairness in prediction improves as quantified by lower values of $SPD$ measure for gender. **B:** Same as **A** but for race. **C:** Fairness implies lower value of $CV_{prob}$ for the model as compared to true data. $CV_{prob}$ for three rating labels lie above the identity line indicating bias in prediction when $HEM$ is not incorporated in the loss function. **D:** Improvement in fairness w.r.t $CV_{prob}$ for all 6 rating labels for model with $HEM$ in the loss function.

the loss defined in (1). The code can be found here [3].

**-Remark:** Note that, the loss function does not have any explicit component for reducing unfairness w.r.t sensitive attributes (race and gender). Still minimizing this loss function improves fairness as we can see in the next section. This verifies that $HEM$ metric indeed is a quantification of quality of a talk which is affected by unfair rating.

**-Results:** We have used a feed forward neural network with one hidden layer for classification. The hidden layer consists of 400 hidden nodes. From (1) we can see that in case of $\epsilon = \lambda = 0$ the model does not use the knowledge of $HEM$ metric, hence it only focuses on accuracy. With the increase of $\epsilon$ and $\lambda$ it aims to improve fairness which leads to reduction in model performance as a trade-off. So one big challenge is to find a good set of $\epsilon$ and $\lambda$ which optimizes both accuracy and fairness. We performed a grid search to find the suitable hyperparameters. For $\epsilon$ we used the range from $0.01$ to $0.2$ and $\lambda$ lies between $0.2$ and $10$. Results for relevant choice of $\epsilon$ and $\lambda$ are reported in Table 3 in Appendix. We use binary accuracy (percentage of data correctly classified) to measure the performance of the model and $SPD$ [7] to quantify the fairness, where $SPD$ measures the difference of rating predictions between two groups. Formally, SPD $= |\mathbb{P}(y_i = 1|S \in G_1) - \mathbb{P}(y_i = 1|S \in G_2)|$ We measure $SPD$ both w.r.t. gender (i.e. $G_1 =$ Male and $G_2$ = other gender) and race (i.e. $G_1 =$ White and $G_2$ = other race). Smaller values of SPD indicates better fairness.

The average accuracy of all 6 rating labels is 67% when we train our model without any $HEM$ loss ($\epsilon = \lambda = 0$). After

[3] https://github.com/racharyy/TED_HEM

Table 3: Results of TEDtalk rating prediction. Accuracy is the binary accuracy of the classification model. $\epsilon$ and $\lambda$ are the hyperparameters as defined in (1). M-SPD (Gender) denotes SPD of the predicted label w.r.t. gender. M-SPD (Race) similarly defined for race. True-SPD (Gender) denotes the SPD of the true dataset w.r.t. gender. Lower values of SPD implies better fairness. It is to see that SPD of the predicted labels is best when the model is trained using HEM loss. Best (lowest) SPD w.r.t. race is achieved across all ratings when trained with HEM. In case of *longwinded* rating the SPD w.r.t. gender has increased from true SPD. However, this is much lower than the M-SPD when trained without HEM loss.

| Ratings | $(\epsilon, \lambda)$ | Accuracy ↑ | M-SPD (Gender) ↓ | True SPD (Gender) | M-SPD (Race) ↓ | True SPD (Race) |
|---|---|---|---|---|---|---|
| fascinating | (0,0) | 0.77 | 0.18 | 0.16 | 0.33 | 0.38 |
| | (0.02,4) | 0.77 | 0.16 | | 0.30 | |
| | (0.017,5) | 0.75 | **0.06** | | **0.19** | |
| ingenious | (0,0) | 0.76 | 0.27 | 0.18 | 0.18 | 0.17 |
| | (0.02,4) | 0.72 | 0.15 | | **0.11** | |
| | (0.017,5) | 0.73 | **0.15** | | 0.29 | |
| jaw-dropping | (0,0) | 0.69 | 0.21 | 0.15 | 0.09 | 0.08 |
| | (0.02,4) | 0.65 | **0.09** | | **0.004** | |
| | (0.017,5) | 0.61 | 0.13 | | 0.18 | |
| longwinded | (0,0) | 0.61 | 0.22 | 0.01 | 0.12 | 0.14 |
| | (0.02,4) | 0.55 | **0.12** | | **0.04** | |
| | (0.017,5) | 0.59 | 0.17 | | 0.17 | |
| unconvincing | (0,0) | 0.61 | 0.05 | 0.06 | 0.15 | 0.30 |
| | (0.02,4) | 0.59 | 0.07 | | **0.01** | |
| | (0.017,5) | 0.57 | **0.02** | | 0.02 | |
| ok | (0,0) | 0.61 | 0.06 | 0.11 | 0.18 | 0.10 |
| | (0.02,4) | 0.58 | 0.01 | | **0.001** | |
| | (0.017,5) | 0.61 | **0.01** | | 0.13 | |

adding the $HEM$ loss to our model (with $\epsilon = 0.017$ and $\lambda = 5.0$), we achieve 64% accuracy on rating prediction but the fairness w.r.t. gender improved significantly as shown in Figure 7 A. For the choice of $\epsilon = 0.02$ and $\lambda = 4.0$ we observe similar model performance (64% accuracy) and fairness w.r.t. race (Figure 7 B). Note that, the model's accuracy is decreased by 3% only, when trained with $HEM$ in the loss function. The goal of this work is not to beat state of the art in prediction accuracy but to achieve improved fairness in prediction for comparable accuracy levels.
In general, the $SPD$ metric quantifies fairness in prediction between two complementary groups (e.g. *male* and *not male*). In our dataset, there are 3 gender labels and 4 race labels, allowing us to measure fairness of the model's prediction across 12 different pairs of groups. Previous work on fairness in public speech introduced a novel metric, $CV_{prob}$, to collectively quantify fairness across all these 12 groups by measuring their variability [2]. Intuitively, $CV_{prob}$ compares variability of ratings across possible instances of sensitive attributes (here, race and gender) for the prediction model with and without incorporation of fairness. For example, let us assume that White and African American speakers have probabilities of 0.6 and 0.3 to be rated *fascinating*. It would be expected that after incorporation of fairness into the prediction model this variability in rating probabilities will drop, becoming 0.58 and 0.55 say. Hence lower the variability of rating probabilities across 12 groups as compared to true data, lower will be the values of $CV_{prob}$ for the fair model and higher will be the fairness in prediction (details about the metric can be found in Section 7.3 of [2]). We show in Figure 7 (C and D) that our model when trained with $HEM$ loss reduces the variability in rating probability across 12 groups for each of the 6 ratings considered (all dots lie under the identity line) as compared to true data. It is important to note that $CV_{prob}$ for model with $HEM$ in loss function not only lies below identity line but has lower magnitude for all 6 ratings as compared to $CV_{prob}$ for model without $HEM$ in loss function (compare y-axis values for C and D in Figure 7), indicating increase in fairness across all rating labels.

# 8 Conclusion

In this work we take the following steps to build a fair rating prediction model for public speeches: 1) we first define a heterogeneity metric $HEM$, that quantifies the quality of a talk based on variation of characteristics in transcript and gestures used by a speaker, 2) we identify a meaningful representation of heterogeneity in both verbal and non verbal domain of public speeches to define $HEM_{tr}$ and $HEM_{ges}$ 3) we establish credibility of $HEM$ by showing that it has meaningful relationship with viewer ratings for TED talks for both positive and negative rating labels 4) we also find that $HEM$ successfully captures the bias in ratings for public speeches 5) finally, we define a meaningful loss function by incorporating $HEM$ to design a fair prediction model. This work therefore combines fairness in the rating of public speeches with the quality of a speech based on heterogeneity in verbal and non-verbal aspects. It also brings together the use of an intuitive loss function with the computational power of a neural network. Though our focus was mostly on verbal and non-verbal aspects of the speeches, our work can easily be generalized to other modes. The simplicity of our encoding systems for defining $HEM$ also makes it applicable to other similar datasets. Besides these, information from temporal evolution can also be used to improve our model in future work.

# References

[1] H Porter Abbott. *The Cambridge introduction to narrative*. Cambridge University Press, 2008.

[2] Rupam Acharyya, Shouman Das, Ankani Chattoraj, and Md Iftekhar Tanveer. Fairyted: A fair rating predictor for ted talk data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 338–345, 2020.

[3] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 38–45, 2019.

[4] William Ball. *A sense of direction: Some observations on the art of directing*. Quite Specific Media Group, 1984.

[5] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.

[6] Harold Barrett. The sophists rhetoric, democracy, and plato's idea of sophistry. 1987.

[7] Dan Biddle. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, Ltd., 2006.

[8] Lloyd F Bitzer. The rhetorical situation. *Human Communication: Core Readings*, pages 303–16, 1999.

[9] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[10] Steven Robert Brydon and Michael D Scott. *Between one and many: The art and science of public speaking*. Citeseer, 2003.

[11] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[12] Tony Carlson. *The how of wow: A guide to giving a speech that will positively blow'em away*. Amacom Books, 2005.

[13] Lei Chen, Ru Zhao, Chee Wee Leong, Blair Lehman, Gary Feng, and Mohammed Ehsan Hoque. Automated video interview judgment on a large-sized corpus collected online. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 504–509. IEEE, 2017.

[14] Charles S Costello. A psychological approach to public speaking. 1930.

[15] Peter DeCaro, Tyrone Adams, and Bonnie Jefferis. audience analysis. *Preuzeto*, 9:2018, 2012.

[16] Ganga Dhanesh. speaking to a global audience. In *Public Speaking: The Virtual Text. ScholarBank@ NUS Repository*. 2012.

[17] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[18] Rosenberg Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[19] Walter R Fisher. Human communication as narration: Toward a philosophy of reason, value, and action. 1989.

[20] Rick Garlick. Verbal descriptions, communicative encounters and impressions. *Communication Quarterly*, 41(4):394–404, 1993.

[21] Kim Giffin and Bobby R Patton. Fundamentals of interpersonal communication. 1971.

[22] Hamilton Gregory. *Public speaking for college and career*. McGraw-Hill Higher Education, 2010.

[23] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, page 2, 2016.

[24] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6. IEEE, 2009.

[25] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

[26] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

[27] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[28] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[29] Stephen Lucas and Paul Stob. *The art of public speaking*. McGraw-Hill New York, 2004.

[30] Myron W Lustig, Jolene Koester, and Rona Halualani. *Intercultural competence: Interpersonal communication across cultures*. Pearson/A and B, 2006.

[31] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[32] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[33] Albert Mehrabian. Silent messages: Implicit communication of emotions and attitudes, wadsworth pub. *Co.*, 1981.

[34] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[35] AH Monroe and D Ehninger. Principles and types of speech compression. *Glenview, IL: Scott Foreman*, 1974.

[36] 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, 2016.

[37] James W Neuliep. *Intercultural communication: A contextual approach*. SAGE Publications, Incorporated, 2020.

[38] Ann Neville Miller. An exploration of kenyan public speaking patterns with implications for the american introductory public speaking course. *Communication Education*, 51(2):168–182, 2002.

[39] Charles M Newcomb. The educational value of expression. *Quarterly Journal of Speech*, 3(1):69–79, 1917.

[40] Laurent Son Nguyen and Daniel Gatica-Perez. Hirability in the wild: Analysis of online conversational video resumes. *IEEE Transactions on Multimedia*, 18(7):1422–1437, 2016.

[41] Jamie Oliver. Teach every child about food. *TED Award Speech, February*, page 45, 2010.

[42] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[43] Christof Rapp. Aristotle's rhetoric. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2010 edition, 2010.

[44] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[45] Warren Sandmann. introductions & conclusions. *Public Speaking: The Virtual Text*, pages 1–12, 2013.

[46] Juliann C Scholl. special occasion speaking. In *Public speaking: The virtual text*. 2013.

[47] Lisa Schreiber and Morgan Hartranft. introduction to public speaking. 2013.

[48] Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H Chi. Transfer of machine learning fairness across domains. *arXiv preprint arXiv:1906.09688*, 2019.

[49] Herbert A Simon. Models of bounded rationality (volume iii), 1982.

[50] Jo Sprague and Douglas Stuart. Tlie speaker's handbook, 1984.

[51] M Iftekhar Tanveer, Samiha Samrose, Raiyan Abdul Baten, and M Ehsan Hoque. Awe the audience: How the narrative trajectories affect audience perception in public speaking. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 24. ACM, 2018.

[52] Stella Ting-Toomey and Tenzin Dorjee. *Communicating across cultures*. Guilford Publications, 2018.

[53] Nikolaj Tollenaar and PGM Van der Heijden. Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2):565–584, 2013.

[54] Xiaosui Xiao. From the hierarchical ren to egalitarianism: A case of cross-cultural rhetorical mediation. *Quarterly Journal of Speech*, 82(1):38–54, 1996.

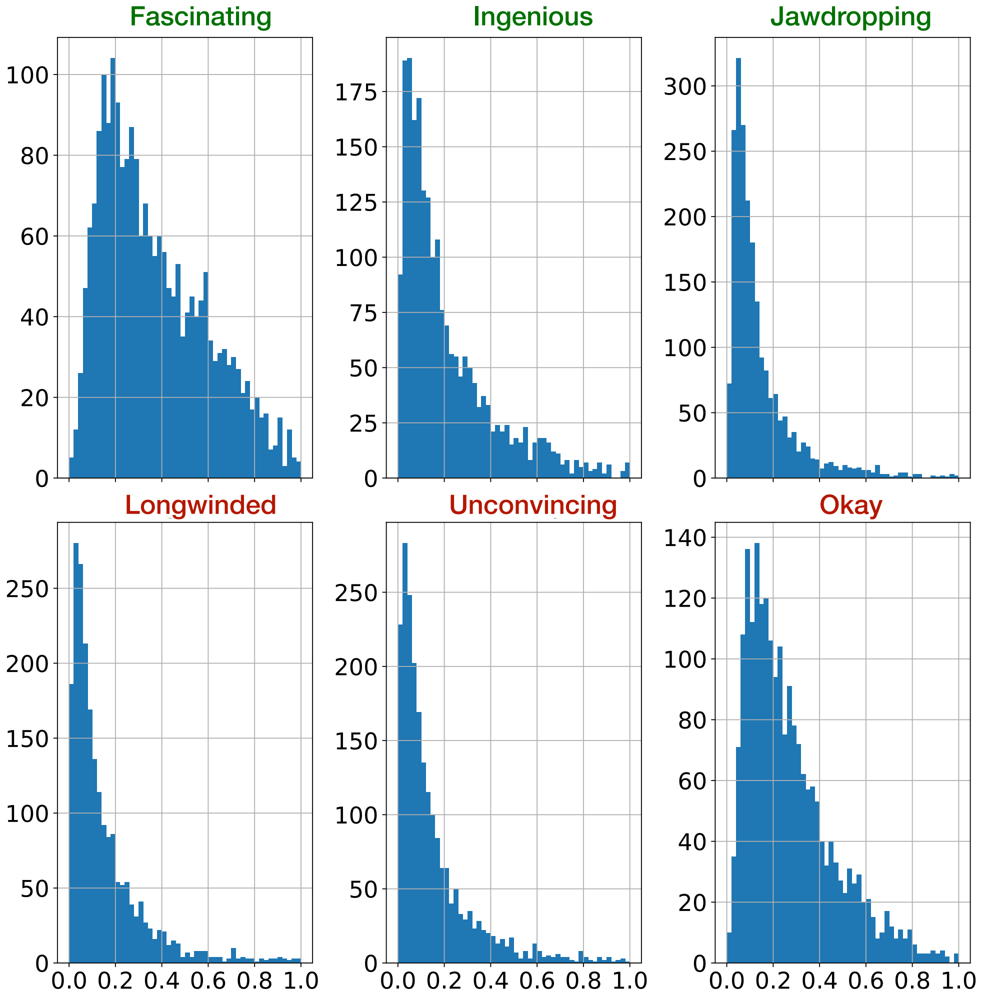

Figure 8: Distribution of non-binarized rating values. Higher mass on smaller values of ratings justifies the small change of rating w.r.t. HEM metric in Figure 3.

# A Extra Figures

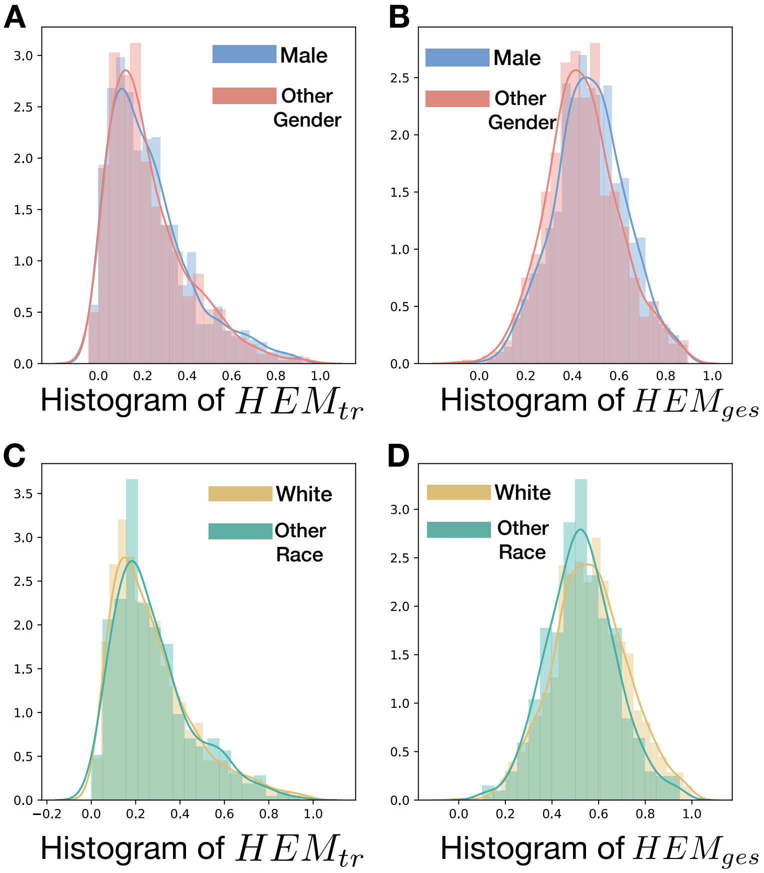

Figure 9: $HEM_{tr}$ and $HEM_{ges}$ are not biased by definition
A-B: Shows histogram of $HEM_{tr}$ and $HEM_{ges}$ for male speakers (blue) and speakers of other gender (red). The almost overlap between the two histograms indicate that the bias observed in Figure 5 and 6 are genuine and not an artifact due to a biased metric. C-D: Same as A-B but for race comparing white speakers (yellow) with speakers of other gender (green) for both $HEM_{tr}$ in C and $HEM_{ges}$ in D. We find no significant difference in the two histograms.

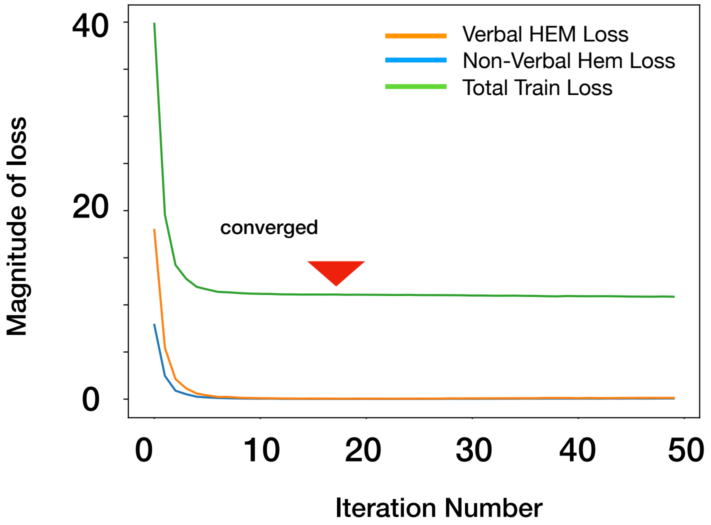

Figure 10: Convergence in training neural network
The loss function decreases and finally converges with increase in iteration steps indicating that learning is complete in the neural network. We show training loss w.r.t $HEM_{tr}$ in orange, $HEM_{ges}$ in blue and total training loss in green.

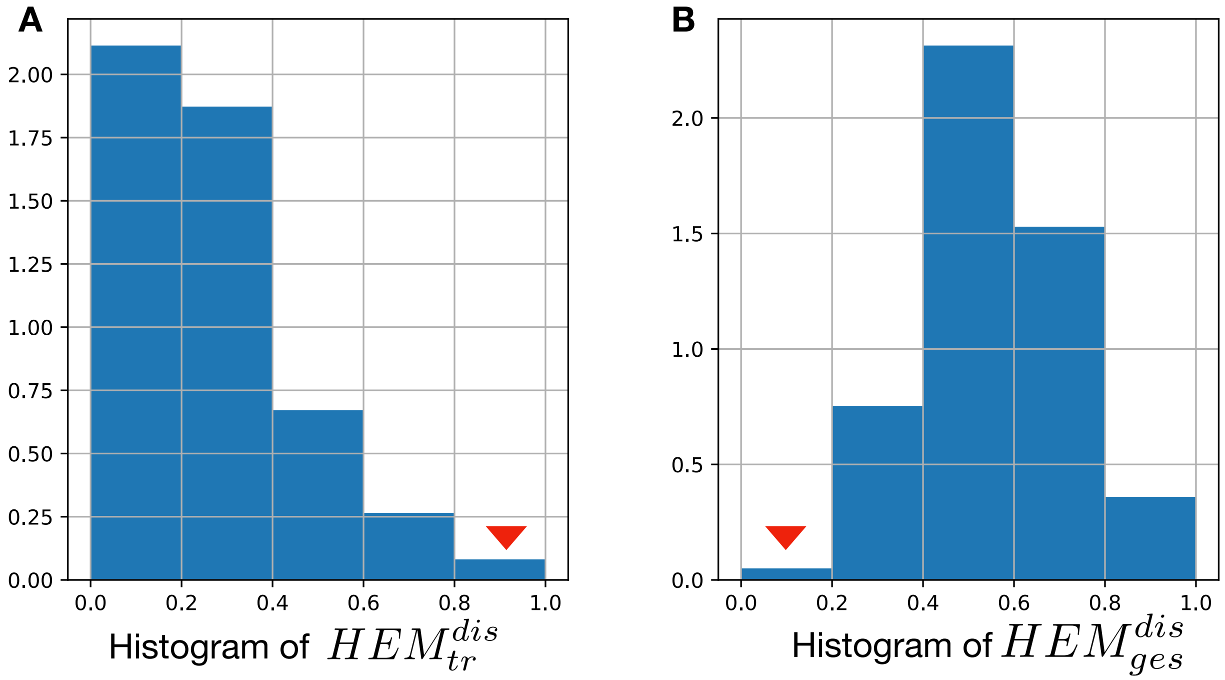

Figure 11: Distribution of discretized $HEM_{tr}^{dis}$ and $HEM_{ges}^{dis}$
A: $HEM_{tr}$ is discretized by assigning values between $[0.0, 0.2)$ to category $0$, $[0.2, 0.4)$ to category $1$ and so on $[0.8, 1.0]$ to $4$. We find negligible instances for $HEM_{tr}^{dis} = 4$, (i.e, $HEM_{tr} >= 0.8$ and $<= 1.0$) as shown in red. B: Similarly, $HEM_{ges}^{dis} = 0$, (i.e, $HEM_{ges} >= 0.0$ and $< 0.2$) is almost non-existent, pointed by red arrow.