Multimodal, Multi-Class Bias Mitigation for Predicting Speaker Confidence

Andrew Emerson
Educational Testing Service
aemerson@ets.org

Arti Ramesh
Nissan Advanced Technology Center
artiramesh01@gmail.com

Patrick Houghton
Educational Testing Service
phoughton@ets.org

Vinay Basheerabad
Educational Testing Service
vbasheerabad@ets.org

Navaneeth Jawahar
Educational Testing Service
njawahar@ets.org

Chee Wee Leong
Educational Testing Service
cleong@ets.org

ABSTRACT

Projecting confidence during conversation or presentation is a critical skill. To effectively display confidence, speakers must employ a blend of verbal and non-verbal signals. A predictive model that leverages rich multimodal cues to measure a speaker's confidence must also mitigate biases that develop through data labelling practices, inherent imbalances in the demographic distribution, or biases introduced into the model during the training process. Fairly predicting the confidence of speakers across differing backgrounds enables more accurate and actionable feedback to a larger population of speakers. This paper introduces a set of approaches for bias mitigation for multimodal, multi-class confidence prediction of adult speakers in a work-like setting. We evaluate the extent to which bias mitigation techniques improve the performance of a multimodal confidence classifier with a dataset of 233 2-minute videos. Experimental results suggest that by bounding the loss across perceived races, genders, accents, and ages, multimodal models can significantly outperform unmitigated baselines. The implications, including automated feedback of speaker confidence, are discussed.

Keywords

Bias Mitigation, Multimodal Learning Analytics, Confidence Measurement.

1. INTRODUCTION

Confidence is an essential component of effective communication in the 21st century, belonging to a larger set of transferrable skills that are critical for success in the workplace [20]. Projecting ample confidence in communication involves a multitude of behaviors, including both non-verbal and verbal cues [9] [15]. For speakers who struggle with this skill, practicing can help prepare them for high-stakes scenarios, such as a job interview or a class presentation. Both students and educators require this skill for clear and effective communication, and fortunately, speakers may improve their projected confidence through targeted learning modules or practice environments. A system that allows speakers to practice their communication skills must be able to classify the speaker's level of projected confidence accurately and *fairly* to ultimately give insights for improvement. However, studies have shown that prevailing technologies in this context struggle with the complexity of the confidence construct and with the collection, diversity, and fairness of training data [14], [17], [22] [24]. The resulting models may be biased against subpopulations of speakers that are underrepresented in the training data, and the algorithms used to model confidence may magnify this bias without intentional mitigation efforts.

To train computational models of speaker confidence, we must carefully consider the demographic makeup of the data for training. Cultural, gender, racial, and age differences may not only affect how a person may project confidence, but these differences can also affect how a listener or observer *perceives* the speaker's confidence. A fair, systematic annotation system to label speakers' confidence levels is a necessary first step to overcome this issue, but it is not foolproof. Biases exist in annotations lacking in diversity among labelers. Beyond labels, the data itself must be expressive enough to convey the intricacies and nuance of the speaker's communication style. As such, multimodal data that includes both the verbal (e.g., speaking rate) and visual (e.g., eye contact or facial expressions) can aid in the computational representation of confidence. Models that leverage the crafted labels and rich multimodal data must then map unseen videos of speakers to accurate confidence labels without discriminating unfairly against groups that were overlooked in the training process. For high-stakes application of a perceived confidence measurement system (e.g., video interview assessment) it is important that we perform bias mitigation of the underlying model to inspire trust and increase utility across a wide range of different use cases.

In this paper, we introduce a framework to predict speaker confidence in 2-minute video presentations by leveraging multimodal data. Importantly, we establish a pipeline for integrating bias mitigation algorithms from the FairLearn [4] open-source library to overcome disparities in the confidence labels between subpopulations of the speaker data. The subsequent analyses and models use data collected from a set of 233 videos that were annotated for confidence using a thorough, systematic labelling process. We conduct a series of experiments to evaluate the use of bias mitigation techniques for four perceived demographic splits of the speaker data: gender, race, accent, and age. In each perceived demographic split, we show that bias mitigation yields improved results over baselines that do not mitigate biases. We discuss insights from our investigations and present practical considerations involved in mitigating bias.

A. Emerson, A. Ramesh, P. Houghton, V. Basheerabad, N. Jawahar, and C. W. Leong. Multimodal, multi-class bias mitigation for predicting speaker confidence. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 610–616, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729896

2. RELATED WORK

2.1 Perceived Confidence

Using vocal and linguistic cues, [15] explored how expressed confidence and doubt in speech influence perceptions of trustworthiness and persuasiveness. The experiments they conducted concluded that listeners' ratings of speaker confidence correlated with the intended confidence level, affected by the statement's communicative function and specific introductory phrases. Additional acoustic analysis identified patterns in pitch, intensity, and speech rate that vary with perceived confidence levels, offering insights into how vocal and linguistic signals convey metacognitive states like certainty or uncertainty to listeners. In an ensuing study [19], the authors found that observers can discern a speaker's confidence level based on visual cues, particularly in situations where speech content confidence varies. By analyzing speakers' facial expressions, eye movements, and head movements in videos without sound, observers were able to accurately gauge speakers' confidence levels during general knowledge question responses. These findings corroborate conclusions in an earlier study by Walker [23], who investigated the impact of verbal and nonverbal cues on perceptions of confidence, using recordings of actresses with varying levels of confidence shown through both cues. The findings revealed that nonverbal cues significantly outweighed verbal cues in influencing the audience's impressions, which is aligned with prior research suggesting nonverbal signals are more influential than verbal signals in conveying emotions and feelings.

2.2 Bias Mitigation and Fairness Metrics

Gupta et al. introduce a way to detect and mitigate bias in an authentic educational setting by using a deep learning-based stealth assessment framework designed for game-based learning environments [11]. The authors aimed to predict students' reflection depth and post-test science scores by analyzing written reflections and game interactions. It addresses the fairness of the resulting predictive models by investigating and mitigating biases related to gender and prior gaming experience, using the ABROCA statistic for bias measurement. The research, conducted with 119 students in a microbiology game-based learning setting, confirms the models' effectiveness and the impact of debiasing techniques in achieving fair assessment outcomes. In another line of work related to fairness and biases in education, the authors identified causes, the groups most affected, and the stages in educational algorithm development and deployment where bias can arise [2]. The authors review existing empirical evidence on bias, particularly focusing on race/ethnicity, gender, nationality, and they extend to less-studied categories like socioeconomic status and disability, proposing a shift from identifying unknown biases to achieving equity. The paper also outlines obstacles to overcoming algorithmic bias and suggests key areas for effort to mitigate these biases in Artificial Intelligence in Education (AIED) systems and educational technologies. Similarly, in terms of high-stakes educational assessment, [18] addresses the challenge of algorithmic fairness and bias, particularly in educational tools utilizing NLP and speech processing technologies, which may inadvertently perpetuate biases against certain groups. The authors explore various definitions of fairness and their application in educational contexts, using data to show how biases, such as those based on native language backgrounds, can affect scores in English proficiency assessments. The discussion acknowledges the complexity of achieving total fairness, suggesting that different interpretations of fairness may necessitate distinct approaches. More generally, [16] examines how the adoption of predictive models in education raises fairness concerns, echoing past issues of bias in educational access. The authors emphasize the need for policymakers and developers to address biases proactively to promote equitable educational outcomes. Additionally, [3] discusses racial bias in predicting student success. Specifically, this study explores algorithmic bias in higher education predictive models for course and degree completion, showing that such bias could reduce support for marginal Black students. The severity of bias varies, being notably higher for students predicted to be most at risk. The findings highlight the importance of context in addressing algorithmic bias and suggest that enhancing data collection on Black students could mitigate bias, indicating a need for more nuanced data to improve prediction accuracy and equity.

As far as the workplace is concerned, [13] examines how descriptive and prescriptive gender stereotypes in the workplace hinder women's career advancement by leading to biased judgments and decisions. The authors explain that descriptive stereotypes contribute to gender bias through negative expectations about women's fit for male-dominated roles, while prescriptive stereotypes create behavioral norms that penalize women for deviations or inferred violations, especially in success scenarios. The research explores the career impacts of such biases and the conditions that amplify or mitigate their effects. Often compounded with gender bias is race bias. Survey results by [8] show that as women and minorities climb the workplace ladder, they face more inequality compared to white men, with black women specifically suffering from direct discrimination. White men tend to gain power through similar networks, indicating different reasons for the power gap across groups, suggesting the need for varied solutions. Another notable workplace bias relates to age. Finkelstein et al. [10] utilize the tripartite model of attitudes and Fiske's social bias framework to discuss research, motives, and impacts of age bias, alongside practical strategies for reduction from various perspectives, concluding with future research directions. Even spoken accent can be a discriminant. A study conducted by Deprez-Sims and Morris [7] reveals how accents affect job applicant evaluations in the U.S., showing a preference for Midwestern accents over French, with Colombian accents judged neutrally, suggesting bias is mediated by perceived similarity, underscoring the importance of auditory cues in employment discrimination research. Our paper builds upon these frameworks by analyzing bias for the four demographic groups: gender, race, age, and accent.

3. VIDEO DATASET

This study investigates bias mitigation using a dataset of speakers who recorded videos of themselves in a virtual practice interview setting [5]. Participants recorded themselves for up to two minutes using their computer's webcam and microphone within the browser-based application. Each speaker was asked to answer up to eight total prompts related to common interview questions. Participants were recruited via Amazon Mechanical Turk throughout the United States. A total of 260 speakers recorded themselves for a total of 1891 videos (7.27 prompts answered per speaker). All participants in this study provided written consent.

3.1 Confidence Annotations

To develop an automated model of speaker projected confidence, ground truth labels of projected confidence were first annotated. Partnering with Scale.ai, we selected a subset of the data to be labeled for confidence to capture ratings from unique speakers for a single prompt (Table 2). The subset consisted of 233 total 2-minute videos. The confidence rating, of either low, medium, or high confidence was first developed in-house using a rubric validated by psychologists and subject matter experts (Table 1).

Table 1. The rubric used to annotate 233 2-minute videos for confidence.

Confidence Level (Label)	Eye Gaze	Gestures and Body Movement	Posture	Vocal Variation	Facial Expression	Speaking Pace
Low (0)	Inconsistent eye gaze with frequent changes in focal point, such as downward or upward gaze.	Consistent distracting movements, such as fidgeting or extraneous body movement.	Consistently poor posture that is closed or "protective" and/or tense.	Lack of variation in tone.	Constant changes in facial expressions that detract from what is being said.	Hesitation between and within sentences; self-correction and/or false starts.
Medium (1)	Eye gaze toward camera with some inconsistencies and changes in focal point.	Some inconsistent gesturing and body movement that are misaligned with what's being said, distracting movements such as fidgeting.	Some inconsistent posture, such as slouching or tenseness at times.	Some inconsistent variation in tone, but still comes across as somewhat comfortable.	Some inconsistent facial expressions that are misaligned from what is being said.	Some inconsistent pacing; some hesitation at times.
High (2)	Consistent eye gaze toward camera.	Intentional gesturing and/or body movement that aligns with what is being said.	Maintains consistent, attentive, and upright posture that appears natural.	Appropriate variation in tone that is relaxed or projected (as appropriate for a presentation).	Facial expression appears comfortable and natural.	Consistent and steady pacing; does not focus on errors (correct and/or move on).

The rubric was refined from an original four-point scale that yielded extremely skewed distributions to the current three-point variation. The annotation process required several iterations to fine-tune the rubric such that Scale annotators were able to independently achieve consistent ratings that aligned with our team's evaluation ratings. Each iteration was conducted with a very small subset of the videos that were labelled by our team for comparison. Once completed, Scale conducted an annotation process to label each video, with 10 annotations per video. Consequently, we received a finalized annotation label per video accounting for reliability of the 10 annotators using a proprietary averaging technique by Scale. The distribution of annotations for this dataset of 233 videos included 49 (21%) low confidence, 143 (61%) medium confidence, and 41 (18%) high confidence.

Table 2. The prompt used for the 233 speakers.

Prompt
Please tell us about a work situation in which you were not the formal leader but tried to assume a leadership role. Please provide details about the background of the situation, the behaviors you carried out in response to that situation, and what the outcome was.

3.2 Demographic Annotations

In addition to the annotations for confidence, the videos were also labeled to acquire *perceived demographic* information about the speakers. For each unique speaker, five annotators from the internal team rated four relevant demographic areas to this study: gender (male or female), race (White, Black or African American, Asian or Asian American, Hispanic or Latino, or Other), age (18 through 25, 26 through 35, 36 through 45, 46 through 55, 56 through 65, or 66 and older), and accent (native accent or non-native English-speaking accent). A final rating for each demographic characteristic was achieved by either averaging the five ratings or by taking the most common annotation, depending on the type of demographic (e.g., taking the average perceived age, taking the most annotated race). The demographic annotations for the subset of speakers who were annotated for confidence are shown in Table 3. Each of these characteristics is the annotators' perception of the demographic subgroup. As such, the speaker may identify differently, but the original study in which these videos were collected did not include a demographic questionnaire to allow speakers to report these data. As they relate to bias and fairness, both perceived and actual demographic characteristics play a heavy role [6]. This study focuses on the perceived gender, race, age, and accent of the speakers, consistent with perceived confidence construct we are measuring.

Table 3. Demographic annotations for the 233 2-minute video dataset.

Perceived Demographic		Count (%)
Gender	*Male*	109 (47%)
	Female	124 (53%)
Race	*White*	159 (68%)
	Non-White	74 (32%)
Age	*Under 35*	138 (59%)
	Over 35	95 (41%)
Accent	*Native*	207 (89%)
	Non-Native	26 (11%)

For the purposes and scope of this investigation, we aggregated the demographic data in several ways to account for the unequal distributions of certain demographic characteristics in the dataset and to evaluate frequently occurring biases in other applications. First, due to skewed numbers of racial data (159 White, 51 Black or African American, 14 Asian or Asian American, 8 Hispanic or Latino, and 1 Other), we chose to split the data into binary categories of White (*n*=159) and Non-White (*n*=74). Second, again due to distribution considerations for age groups (29 18 through 25, 109 26 through 35, 54 36 through 45, 31 46 through 55, 10 56 through 65, and 0 66 and older), we aggregated the age demographic into two binary categories, under 35 (*n*=138) and over 35 (*n*=95).

3.3 Demographic Distributions

A key factor in understanding and detecting inherent biases in this dataset is the distribution of labels (i.e., confidence) per demographic. If we detect disparities or skewed distributions of confidence labels across demographics, this could have significant, unintended effects downstream in prediction settings. For example, if a new speaker from an affected demographic interacts with a confidence model that was trained from a dataset that included few examples of high confidence for that speaker's demographic, this could result in incorrect predictions. Ideally, predictive models are exposed to a diverse set of data points. In our setting, we investigate the distribution of confidence labels for four demographic splits: (1) gender, (2) race, (3) age, and (4) accent. In Table 4, we illustrate the distributions of confidence labels for each of the four demographic splits mentioned.

Table 4. Perceived demographic distributions for each confidence label.

Perceived Demographic		**Confidence Labels**		
		Low (%)	**Medium (%)**	**High (%)**
Gender	*Male*	29 (27%)	66 (61%)	14 (13%)
	Female	20 (16%)	77 (62%)	27 (22%)
Race	*White*	36 (23%)	93 (58%)	30 (19%)
	Non-White	13 (18%)	50 (68%)	11 (15%)
Age	*Under 35*	34 (25%)	81 (59%)	23 (17%)
	Over 35	15 (16%)	62 (65%)	18 (19%)
Accent	*Native*	45 (22%)	125 (60%)	37 (18%)
	Non-Native	4 (15%)	18 (69%)	4 (15%)

In addition to the distributions of each individual demographic, we examine the distribution of the intersection of two frequently studied demographics: gender and race (Table 5). This intersection results in the following demographic combinations: white and male (*n*=75), white and female (*n*=84), non-white and male (*n*=34), and non-white and female (*n*=40).

Table 5. Confidence label distributions for the intersection of perceived gender and race.

Perceived Demographic	**Confidence Labels**		
	Low (%)	**Medium (%)**	**High (%)**
White, Male	21 (28%)	44 (59%)	10 (13%)
White, Female	15 (18%)	49 (58%)	20 (24%)
Non-White, Male	8 (24%)	22 (65%)	4 (12%)
Non-White, Female	5 (13%)	28 (70%)	7 (18%)

3.4 Multimodal Feature Extraction

Using the visual and speech components of the 233 2-minute videos, we established a multimodal feature extraction pipeline to derive salient features that would model speaker confidence. A total of nine features were extracted from the videos (three from the visual modality and six from the speech modality).

For the visual modality, we extracted features related to the speaker's eye gaze. Specifically, we extracted features that represented anomalies in the speaker's eye movement. These anomalies were computed in three dimensions (horizontal or "left to right", vertical or "up and down", and optical or "to and from the webcam"), resulting in three separate anomaly features. The motivation behind the anomalies was that a speaker will have fewer eye gaze anomalies in any direction when they are more confident. The computational process to derive the anomalies, along with the confirmation of the hypothesis, can be found in a prior study [9].

For the speech modality, we extracted features using our in-house speech processing system that performs both automatic speech recognition and statistical analysis of speech patterns. In total, we extracted six speech features using this system, including the speaker's response length, the number of silences per token, the pause ratio, the number of hesitation markers, the filler ratio, and the speaking rate. A comprehensive definition of these features is provided in a prior study [9].

4. MULTIMODAL CONFIDENCE MODELS

Leveraging the multimodal features from both the speech and visual modalities, we constructed models to classify speaker confidence on a scale of low, medium, or high. We compared two approaches: a model that did not utilize bias mitigation techniques and a model that incorporated bias mitigation techniques. For each approach, we implemented a model to compare the performance for each demographic split. Baseline models were constructed similarly as bias-mitigated models, leaving out all FairLearn mitigation functionality.

4.1 Bias-Mitigated Models and Metrics

To mitigate bias in this dataset and to produce fair, unbiased predictive models that can accurately classify speakers' confidence level, we leveraged the FairLearn open-source library. The Python toolkit offers solutions to analyze data based on sensitive attributes or demographic splits with a variety of metrics. It also serves as a wrapper for scikit-learn-style models, where the user can select an optimization algorithm, constraint, loss function, and provide other parameters to mitigate bias. In our experiments, we explored the use of these mitigation features to construct a pipeline for this dataset. Using the best performing configuration of optimization algorithm, constraint, and loss function, we enhanced our multimodal confidence model and evaluated the performance for each demographic.

A primary reason we chose to use the FairLearn library was the ability to conduct multi-class optimization across sensitive attributes (i.e., demographic splits) with two algorithms: grid search and exponentiated gradient optimization [1]. Both algorithms support the bounded group loss constraint, which forces an upper bound on the loss per user-specified group. To satisfy the constraint, a classifier, h, must meet the criteria below, for each subpopulation, a, of a demographic, A.

$$\mathbb{E}\left[loss\left(Y, h(X)\right)\middle|A=a\right] \leq \zeta \qquad \forall a$$

In our case, the groups are the subpopulations of the training data, which are the demographic splits of gender, race, accent, and age. The upper bound value is customizable based on the prediction type (i.e., regression vs. classification), but the loss function dictates the usefulness. In this paper, we use the mean absolute loss of each subpopulation for the optimization function. As such, the mitigation will target all subpopulations to have a mean loss below the provided upper bound threshold. For predicting speaker confidence, we have labeled the possible perceived confidence values as low, medium, or high, but they are encoded as 0, 1, and 2, respectively. This allows for the optimization to be conducted similarly as it would for a regression problem. For example, a mean loss of

1.0 would mean that the model is rating speakers as an entire category different, on average. This ranking-based classification setting allows for flexibility in the optimization, but other classification formulations would not yield sensible results with a similar upper bound usage. In our experiments, we compare the use of the two optimization algorithms (grid search and exponentiated gradient) and the upper bound thresholds for each subpopulation, and we fix the usage of the bounded group loss constraint and absolute loss. Other mitigation approaches were considered, but these selected algorithms fit the problem statement the best.

To measure the performance of the predictive models, we chose to use standard classification metrics, F1 score and accuracy. There is a large collection of metrics that enable the measurement of fairness in a predictive model, but there are few that handle the computation of multiple output target classes. While FairLearn does currently support multi-class optimization algorithms (e.g., exponentiated gradient), it does not currently support fairness-specific metrics for multi-class outputs, at the time of this writing. Fairness metrics commonly used for binary classification include demographic parity, equalized odds, and equal opportunity, which all aim to capture aspects of the expected value of an input data point, given its sensitive attribute (i.e., demographic value) [1] [12]. The development of these metrics for multi-class outputs is still an ongoing research topic. However, F1 score and accuracy can be leveraged to demonstrate how mitigation techniques improve predictive performance for protected subgroups without sacrificing performance for non-protected subgroups. Should this occur, no subpopulations would incur a performance degradation due to bias.

5. EXPERIMENTAL RESULTS

To compare unmitigated and mitigated multimodal models of speaker confidence, we first developed a baseline model using the XGBoost classifier. We chose this base classifier for its versatility and robustness in tabular supervised settings. It is extremely flexible and configurable, and the research community has leveraged it frequently for its modeling capabilities. For classifying perceived speaker confidence, we fix the hyperparameters for fair comparison between the unmitigated and mitigated configurations. Specifically, we fix the max depth of the candidate trees to be 2, the alpha (i.e., L1 regularization of weights) to 5, and the subsampling of training data to 0.9. Each of these hyperparameters were selected based on prior internal efforts, and they serve to help prevent overfitting of the resulting model.

In our experimentation, we found the grid search optimization algorithm to be inferior to exponentiated gradient in most comparisons. As such, we only report the results using exponentiated gradient algorithm. Using the bounded group loss constraint, we do report the results for two different values of the upper bound of the constraint: 0.25 and 0.5.

We conducted five sets of experiments where we compared mitigation to no use of mitigation. Each set of experiments focused on one demographic subpopulation: gender (Table 6), race (Table 7), age (Table 8), accent (Table 9), and gender intersected with race (Table 10). For model selection, we conducted a variation of stratified cross-validation where the test sets may repeat data points. This approach was selected because we wanted to maintain sufficient data points from underrepresented demographic categories in each train and test fold. When the demographic splits included subpopulations that were not well-represented (e.g., only 26 speakers with a perceived non-native accent), standard stratified cross-validation would result in folds either having no examples of these subpopulations or too few for meaningful distinction. Our sampling approach was accomplished by sampling a pre-set percentage of each demographic subpopulation to construct the train and test sets for each fold. We chose to use 30% as the sampling rate. This means that for a single sampling iteration (i.e., each fold), 30% of each demographic was selected for the test set, and the remaining 70% would be used for training. We repeated this process 10 times and aggregated performance metrics by averaging the results, much like standard cross-validation. To avoid sampling the same data points repeatedly, we used a fixed random seed for all experiments, and for each iteration of the sampling, the random seed was incremented by one to yield a different sample of data. The overall distribution of each subpopulation was maintained, keeping the ratio between the groups intact. In each of the subsequent tables, the results for each column represent a single XGBoost model trained either without mitigation (Unmitigated), mitigation with an upper bound on the loss of 0.25, or mitigation with an upper bound on the loss of 0.5. The results are an average of the 10 iterations of sampling, and the best performing model for each metric and subpopulation is bolded.

Table 6. Bias mitigation for gender.

	Unmitigated	**Mitigated (0.25)**	**Mitigated (0.5)**
Overall F1	0.619	**0.637**	0.615
Male F1	0.661	**0.664**	0.661
Female F1	0.572	**0.606**	0.566
Overall Acc.	0.664	**0.676**	0.664
Male Acc.	**0.695**	**0.695**	**0.695**
Female Acc.	0.638	**0.658**	0.638

Table 7. Bias mitigation for race.

	Unmitigated	**Mitigated (0.25)**	**Mitigated (0.5)**
Overall F1	0.565	**0.614**	0.571
White F1	0.541	**0.598**	0.553
Non-White F1	0.615	**0.645**	0.608
Overall Acc.	0.627	**0.650**	0.632
White Acc.	0.607	**0.633**	0.620
Non-White Acc.	0.671	**0.688**	0.657

Table 8. Bias mitigation for age.

	Unmitigated	**Mitigated (0.25)**	**Mitigated (0.5)**
Overall F1	0.622	0.620	**0.631**
Under 35 F1	0.619	0.624	**0.637**
Over 35 F1	**0.623**	0.611	0.621
Overall Acc.	0.667	0.667	**0.678**
Under 35 Acc.	0.667	0.670	**0.685**
Over 35 Acc.	**0.668**	0.663	**0.668**

Table 9. Bias mitigation for accent.

	Unmitigated	**Mitigated (0.25)**	**Mitigated (0.5)**
Overall F1	0.597	**0.619**	0.608
Native F1	0.601	**0.622**	0.604
Non-Native F1	0.556	**0.583**	0.564
Overall Acc.	0.655	**0.665**	0.655
Native Acc.	0.658	**0.666**	0.656
Non-Native Acc.	0.629	**0.657**	0.643

Table 10. Bias mitigation for the intersection of gender and race.

	Unmitigated	Mitigated (0.25)	Mitigated (0.5)
Overall F1	0.568	**0.604**	0.575
White, Male F1	0.591	0.596	**0.598**
Non-White, Male F1	**0.693**	0.685	0.685
White, Female F1	0.504	**0.558**	0.523
Non-White, Female F1	0.548	**0.631**	0.546
Overall Acc.	0.635	**0.652**	0.636
White, Male Acc.	0.636	0.641	**0.645**
Non-White, Male Acc.	**0.750**	0.730	0.740
White, Female Acc.	0.600	**0.632**	0.604
Non-White, Female Acc.	0.608	**0.650**	0.600

6. DISCUSSION

The mitigation approaches applied to each demographic split yielded improved results when compared to the unmitigated variation. For perceived gender, accent, and race, mitigation when using 0.25 as an upper bound outperforms or performs the same as when not using mitigation. For age, the mitigation with an upper bound of 0.5 outperformed the baseline overall and for speakers who were perceived to be under 35. For speakers perceived to be over 35, mitigation did not appear to improve performance. It should be noted that for age, there did not appear to be a difference between the performance of the two possible subpopulations. When applying mitigation to the intersection of perceived gender and race, we found that both upper bound values produced improvements over the unmitigated baseline.

A surprising result from this set of experiments was the ability of mitigation to improve predictive performance of the model overall on each demographic split. An intuitive expectation for mitigation practices would be that the performance of the model on subpopulations is brought closer for each group, but overall model performance could potentially suffer as a result. In this paper, we demonstrate both an improvement in the performance for all individual subpopulations across each demographic split and an improvement of the model performance overall. As this is a classification prediction, reporting F1 score and accuracy are relevant, especially due to the unequal distribution of labels (Table 4). Both metrics demonstrated the models' improvement. While not as specific as designated metrics for fairness and bias (e.g., demographic parity), the choice of F1 score and accuracy for this analysis has shown that the performance improvements for protected or underrepresented subpopulations did not come with a sacrifice of non-protected group model performance.

A particularly illuminating result was the individual improvements for the combined demographics of gender and race. With the intersection of perceived demographics used, there were fewer data points per subpopulation, forcing the model to rely on fewer examples in training. Speakers who were perceived to be non-white males were predicted with high accuracy (0.750) and F1 (0.693), and by mitigating the model, the performance for this group dropped slightly to 0.730 and 0.685, respectively. However, the performance for several subpopulations that had lower accuracy and F1 increased significantly (white females and non-white females). Notably, the mitigated model performance for non-white females improved the accuracy from 0.608 to 0.650 and the F1 score from 0.548 to 0.631. This improvement in performance, while still not quite to the level of model performance of male subpopulations, is substantial. Each percentage point improvement has real implications. A speaker who receives an incorrect prediction for a construct that lowers their self-efficacy (e.g., being predicted as low confidence when the speaker does not have low confidence) could have the unfortunate effect of lowering the speaker's confidence in future scenarios. For subpopulations that may be more susceptible to machine learning errors, every opportunity for improving model performance should be taken [21].

There are still cases where the model performance of individual subpopulations for a given demographic category is unequal (i.e., room to improve). The starkest differences that remain include the difference in performance when mitigating based on gender, where the F1 score for males is 0.664 and 0.606 for females, using the mitigated model. This difference is an improvement compared to the unmitigated model, but it can be further improved through other bias mitigation techniques. For example, other techniques include preprocessing algorithms that attempt to remove the correlation between sensitive features and non-sensitive features. Other approaches include postprocessing algorithms where additional thresholds are utilized to reach desired fairness metrics.

7. CONCLUSION

Classifying speaker projected confidence has significant potential for supporting speakers who are learning how to communicate effectively. Creating a system that provides fast, personalized insights to a speaker based on their visual and spoken cues enables speakers to become more confident and practiced. Such a system must overcome several challenges, however. First, projected confidence is an inherently multi-faceted construct, and requires the system to leverage synergies between multiple modalities of the speaker's expression of confidence. Second, this construct is highly subjective, and it requires a carefully crafted rubric and annotation process to derive meaningful and accurate labels of the data used to train a model. Third, and a large focus of this paper, the data may not always be distributed in such a way that leads to an unbiased model. Moreover, the mapping between demographics and confidence labels may itself be biased, which leads to a biased model without intervention. In this paper, we began to address these challenges by conducting a thorough annotation process of multimodal speaker data through Scale.ai. Mitigated classification models were trained using this data, resulting in improved model performance overall and for individual demographic groups. Specifically, we evaluated this framework for perceived gender, race, age, accent, and the intersection of gender and race. For future applications, we aim to not only produce unbiased results to speakers, but we aspire to explain the predictive outputs to speakers for transparency and insightful recommendations for improvement.

8. ACKNOWLEDGEMENTS

We would like to thank our colleagues at Educational Testing Service for support with this work. Katrina Roohr and Alison Moore assisted in the annotation and rubric creation process, Rutuja Ubale assisted with speech feature extraction, and Suhasini Kalaiah Lingaiah supported visual feature extraction.

9. REFERENCES

[1] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. 2018. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning* (Stockholm, Sweden). PMLR, 80, 2018, 60-69.

[2] Baker, R. S., and Hawn, A. 2021. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 1-41.

[3] Bird, K. A., Castleman, B. L., and Song, Y. 2024. Are algorithms biased in education? Exploring racial bias in predicting community college student success. *Journal of Policy Analysis and Management*, 1-24. DOI= https://doi.org/10.26300/yd7z-6e20.

[4] Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep.* MSR-TR-2020-32.

[5] Chen, L., Zhao, R., Leong, C. W., Lehman, B., Feng, G., and Hoque, M. E. 2017. Automated video interview judgment on a large-sized corpus collected online. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 504-509.

[6] Cohausz, L., Tschalzev, A., Bartelt, C., and Stuckenschmidt, H. 2023. Investigating the importance of demographic features for EDM-predictions. International Educational Data Mining Society, 125-136. DOI= https://doi.org/10.5281/zenodo.8115647.

[7] Deprez-Sims, A. S., and Morris, S. B. 2010. Accents in the workplace: Their effects during a job interview. *International Journal of Psychology, 45*(6), 417-426.

[8] Elliott, J. R., and Smith, R. A. 2004. Race, gender, and workplace power. *American Sociological Review, 69(3)*, 365-386.

[9] Emerson, A., Houghton, P., Chen, K., Basheerabad, V., Ubale, R., & Leong, C. W. 2022. Predicting user confidence in video recordings with spatio-temporal multimodal analytics. In *Companion Publication of the 2022 International Conference on Multimodal Interaction*, 98-104. DOI= https://doi.org/10.1145/3536220.3558007.

[10] Finkelstein, L. M., Hanrahan, E. A., and Thomas, C. L. 2018. An expanded view of age bias in the workplace. In *Aging and Work in the 21st Century*. Routledge, 59-101.

[11] Gupta, A., Carpenter, D., Min, W., Rowe, J., Azevedo, R., and Lester, J. 2023. Detecting and mitigating encoded bias in deep learning-based stealth assessment models for reflection-enriched game-based learning environments. *International Journal of Artificial Intelligence in Education*, 1-28.

[12] Hardt, M., Price, E., and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Advances in Neural Information Processing Systems*, *29*, 1-9. DOI= https://doi.org/10.48550/arXiv.1610.02413.

[13] Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior, 32*, 113-135.

[14] Hu, Q., and Rangwala, H. 2020. Towards fair educational data mining: A case study on detecting at-risk students. In *Proceedings of the 13th International Conference on Educational Data Mining*, 431-437.

[15] Jiang, X., and Pell, M. D. 2017. The sound of confidence and doubt. *Speech Communication 88* (2017): 106-126.

[16] Kizilcec, R. F., and Lee, H. 2022. Algorithmic fairness in education. In *The Ethics of Artificial Intelligence in Education*. Routledge, 174-202. DOI= https://doi.org/10.48550/arXiv.2007.05443.

[17] Leong, C. W., Roohr, K., Ramanarayanan, V., Martin-Raugh, M. P., Kell, H., Ubale, R., Qian, Y., Mladineo, Z., and McCulla, L. 2019. To trust, or not to trust? A study of human bias in automated video interview assessments. DOI= https://doi.org/10.48550/arXiv.1911.13248.

[18] Loukina, A., Madnani, N., and Zechner, K. 2019. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 1-10. DOI= https://doi.org/10.18653/v1/W19-4401.

[19] Mori, Y., and Pell, M. D. 2019. The look of (un) confidence: visual markers for inferring speaker confidence in speech. *Frontiers in Communication 4* (2019): 63.

[20] Rios, J. A., Ling, G., Pugh, R., Becker, D., and Bacall, A. 2020. Identifying critical 21st-century skills for workplace success: A content analysis of job advertisements. *Educational Researcher, 49*(2), 80-89.

[21] Stinar, F., and Bosch, N. 2022. Algorithmic unfairness mitigation in student models: When fairer methods lead to unintended results. In *Proceedings of the 15th International Conference on Educational Data Mining*, 606-611.

[22] Verger, M., Lallé, S., Bouchet, F., and Luengo, V. 2023. Is your model "MADD"? A novel metric to evaluate algorithmic fairness for predictive student models. In *Proceedings of the 16th International Conference on Educational Data Mining*, 91-102. DOI= https://doi.org/10.5281/zenodo.8115786.

[23] Walker, M. B. 1977. The relative importance of verbal and nonverbal cues in the expression of confidence. *Australian Journal of Psychology, 29*(1), 45-57.

[24] Yan, S., Huang, D., and Soleymani, M. 2020. Mitigating biases in multimodal personality assessment. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, 361-369. DOI= https://doi.org/10.1145/3382507.3418889.