Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.Doi Number

Audio Bias Score: Bias Detection Metric for Gender Bias Detection in Audio Datasets

First A. Author1, Fellow, IEEE, Second B. Author2, and Third C. Author, Jr.3, Member, IEEE

1National Institute of Standards and Technology, Boulder, CO 80305 USA

2Department of Physics, Colorado State University, Fort Collins, CO 80523 USA

3Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA

Corresponding author: First A. Author (e-mail: author@ boulder.nist.gov).

This paragraph of the first footnote will contain support information, including sponsor and financial support acknowledgment. For example,   
“This work was supported in part by the U.S. Depart­ment of Com­merce under Grant BS123456.”

ABSTRACT With the increasing integration of audio-based AI systems into high-stakes domains such as healthcare, law enforcement, and social media, ensuring fairness, particularly in terms of gender bias has become essential. While prior research predominantly addresses disparities in model performance, limited attention has been given to quantifying bias inherent in training datasets. To bridge this gap, this work introduces the Audio Bias Score, a novel metric designed to evaluate gender bias in audio datasets using raw audio features: pitch, energy, amplitude, and voice activity. The proposed metric is intentionally language- and demography-agnostic, avoiding dependencies on speaker attributes such as race, age, or language. The score is derived through polynomial regression with L2 regularization (Ridge regression), ensuring robust and generalizable results across diverse datasets. It spans a range from -10 to 10, where 0 denotes a balanced dataset, 10 indicates complete bias toward male speakers, and -10 reflects complete bias toward female speakers. Evaluation across multiple datasets demonstrated strong predictive performance, with R² values ranging from 0.95 to 0.99. By focusing on dataset-level bias rather than downstream model disparities, the Audio Bias Score provides a scalable and interpretable tool for promoting fairness in audio-based AI systems.

INDEX TERMS gender bias, audio data, artificial intelligence, responsible AI

1. INTRODUCTION

The rapid advancement of Artificial Intelligence (AI) has led to its increasing application in critical decision-making domains. AI systems are trained using various data modalities, one of the most prominent being audio. Beyond the use of audio in voice assistants, audio-based AI models are now deployed in high-stakes applications such as predicting neurodegenerative diseases like Parkinson’s [1], Alzheimer’s [2], and related conditions [3] [4], as well as in search and rescue operations [5] and forensic voice analysis within law enforcement [6] .

As these technologies are integrated into vital sectors, the imperative for fairness and accountability becomes paramount. AI systems must be designed and evaluated to ensure they do not perpetuate societal biases. According to PricewaterhouseCoopers (PwC), an AI system is considered biased when it makes decisions that are systematically unfair to certain groups of people [7]. Such bias can arise from unrepresentative training data that embeds existing social prejudices [8], flawed algorithmic design [7], or real-world interactions that reinforce disparities [8]. Bias in audio-based AI systems, particularly in automatic speech recognition (ASR), has been a growing concern. For instance, the study “Racial disparities in automated speech recognition” by Koenecke et al. (2020) found that leading commercial ASR systems had significantly higher word error rates when transcribing African American Vernacular English (AAVE) compared to Standard American English. This disparity illustrates how underrepresentation and linguistic variation in training data can lead to performance gaps, disproportionately affecting certain racial groups [9].

Similarly, ASR systems integrated into social media platforms such as Facebook, Instagram, and YouTube have shown performance disparities across genders. In the study *“*Twists, Humps, and Pebbles: Multilingual Speech Recognition Models Exhibit Gender Performance Gaps*”* by Attanasio et al. (2024), it was observed that multilingual models such as SeamlessM4T (developed by Meta) and Wav2Vec 2.0 consistently performed better on male speakers. These models form the foundation of several ASR systems used in social media applications, further underscoring the real-world implications of gender bias in widely deployed technologies [10].

These real-world cases highlight the importance of identifying and mitigating bias in audio-based AI systems. This research focuses specifically on gender bias introduced through unrepresentative training datasets. The primary objective is to develop a metric that can quantify the extent of gender bias in audio datasets, independent of downstream model performance. By addressing the dataset as a potential source of bias, this work aims to support the creation of fairer and more reliable AI systems in critical domains.

1. RELATED WORKS AND METHODOLOGIES.

Bias measures are used to quantify and detect whether a system or dataset exhibits unfair treatment toward certain demographic groups. Over the years, multiple approaches have been developed to measure bias in audio-based AI systems, primarily focusing on performance-based statistical metrics.

Traditional measures include False Positive Rate (FPR), False Negative Rate (FNR), Equal Error Rate (EER), and Minimum Detection Cost (minCDet) [11]. These metrics quantify decision-making disparities by evaluating a system’s behavior across demographic groups. FPR and FNR respectively capture the rates of incorrectly classified positive and negative instances, while EER identifies the point at which these two rates are equal across groups. The minimum detection cost combines these error rates while weighing false positives and negatives according to application-specific costs.

In addition to these base metrics, more customized bias quantification methods have emerged. These include G2mindiff, G2avg ratio, and G2avg log ratio, which derive from comparisons across demographic groups using performance metrics (like EER, FPR, FNR) [12]. Specifically:

* G2mindiff captures the gap between the base performance of a demographic group and the best-performing group.
* G2avg ratio compares a group’s metric to the average across all groups.
* G2avg log ratio offers a scaled log-based perspective on these performance gaps.
* ​ is the base metric for the demographic group being evaluated
* ​ is the metric of the best-performing group
* is the average across all demographic groups

For speech recognition systems, metrics such as Character Error Rate (CER) [13] and Word Error Rate (WER) [14] [15] are commonly used. CER measures the difference between predicted and actual characters in transcriptions, while WER compares predicted and actual word-level transcriptions.

While these performance-based metrics effectively evaluate model bias, they primarily focus on outcome-level disparities after training. This makes them sensitive to both algorithmic bias and data-induced bias, without isolating the contribution of the dataset itself. As a result, they do not directly assess the inherent bias in the data used to train the models, biases which may stem from demographic imbalances or poor quality recordings.

To address this limitation, Burkhardt et al. (2024) introduced the Nkululeko framework [16], a tool designed to investigate potential bias within audio datasets before model training. Their approach involves:

* Extracting or predicting features such as gender, age, emotional state (valence, arousal), and signal quality (SDR, PESQ, MOS) using pretrained deep learning models.
* Analyzing statistical correlations between these confounding variables and the dataset labels (e.g., depression, dysarthria diagnosis) using metrics such as Cohen’s d and chi-square tests.
* Identifying biases where certain confounders (e.g., gender or arousal) show strong influence on label distribution.

Although this method begins to target dataset-level bias, it still relies on model-derived predictions, which are susceptible to out-of-distribution (OOD) issues and may introduce additional bias due to the algorithms used, particularly when the datasets differ significantly from the model’s training domain.

Recognizing this gap, the metric proposed in this work specifically focuses on detecting inherent gender bias in audio datasets. It takes a model-agnostic approach, relying solely on raw audio features such as pitch, energy, amplitude, and voice activity. Rather than depending on predicted attributes or evaluating post-training model performance, the metric compares distributional characteristics of male and female audio samples to identify imbalances or disparities in how gender is represented acoustically. This enables early-stage bias assessment without requiring labeled task outcomes, making it especially valuable for dataset auditing prior to model development.

In summary, while the majority of existing literature focuses on model performance disparities, and tools like Nkululeko bridge toward label-aware dataset analysis, the proposed metric contributes a novel direction by offering a direct, interpretable, and gender-specific bias score grounded entirely in the structure of the audio data itself.

METHODOLOGY

1. Selection of Features for equation building.

The selection of features for building the Audio Bias Score was based on a review of previous research in audio-based gender classification systems, focusing on the features most used in these studies. The primary features identified were pitch [17], amplitude [18], energy [19], formants, intonations, and Mel-Frequency Cepstral Coefficients (MFCCs) [20] [21] [22] [23].

However, intonations [24] [25] and formants [26] are dependent on the language and can vary among cultures and ethnic groups. Since the Audio Bias Score is designed to be independent of language and race, and focuses solely on detecting gender bias in datasets, these features were excluded from the metric. Additionally, as noted by Bailey et al. (2021), models using raw audio are more robust to gender bias than those based on hand-crafted features, such as mel-spectrograms [27]. Therefore, the Audio Bias Score metric focuses on raw audio-based features.

The number of audio samples and voice activity per gender were also included in the metric. The number of audio samples helps identify class imbalances, while voice activity quantifies the balance between genders. Differences in voice activity can significantly impact the performance of a model trained on an imbalanced dataset.

Gender-based variations naturally exist in factors such as pitch, amplitude, and energy levels. For example, pitch is typically higher for female speakers, while amplitude and energy levels tend to be consistently higher or lower depending on the gender. To ensure that these inherent characteristics do not skew the bias measurement, the standard deviations of pitch, amplitude, and energy levels were used in the metric.

In conclusion, the Audio Bias Score equation incorporates the following features:

* Count of audio samples per gender
* Voice activity per gender
* Standard deviation of amplitude, energy, and pitch per gender

These features were selected to ensure that the metric accurately detects gender bias without being influenced by language or cultural variations.

1. Extracting, Building and Processing the datasets.

To train and build the Audio Bias Score equation, a base dataset had to be created by extracting feature values from multiple audio datasets. The datasets used to build the base dataset included Common Voice [28], LibriSpeech [29], LibriSpeech Multi-lingual [30], TED-LIUM [31], and the AMI Meeting Corpus [32]. These datasets consist of various splits and subsets for different languages, resulting in a total of 420 rows of feature values.

Since existing bias detection metrics did not consider the specific features in the newly developed metric, defining a dependent variable or bias score was challenging. To address this, a controlled data augmentation technique was applied. The augmentation doubled the size of the dataset, and the following procedure was used:

1. For the first half of the dataset, the values for male features were kept constant, while the female features were gradually reduced in increments of 5%.
2. For the second half, the values for female features remained unchanged, while the male features were gradually reduced in increments of 5%.

This technique allowed the introduction of a bias score that reflected the percentage of change applied to the features. The resulting bias scores were represented as values such as 0.0, 0.5, 1.0, and so on, indicating the level of bias in the dataset based on the adjusted feature values.

1. Building and training the equation.

The training dataset was generated by applying a controlled data augmentation technique to the base dataset. Given the objective of developing a predictive equation using linear regression, it was essential to assess whether the dataset met the fundamental assumptions required for this modeling technique. These assumptions include linearity, independence of observations, homoscedasticity, normality of residuals, and the absence of multicollinearity.

However, the results of assumption checks revealed several violations. The Durbin-Watson statistic was calculated as 1.0276, suggesting potential autocorrelation in the residuals, thereby violating the independence assumption. The Breusch-Pagan test confirmed heteroscedasticity, with an LM statistic of 82.087 (p-value = 1.95 × 10⁻¹³) and an F-statistic of 9.741 (p-value = 8.54 × 10⁻¹⁵). White’s test also supported the presence of heteroscedasticity, returning a test statistic of 210.34 with 88 degrees of freedom and a p-value of 5.12 × 10⁻¹². To assess the normality of residuals, the Shapiro-Wilk test produced a test statistic of 0.957 (p-value = 2.41 × 10⁻¹⁰), and the Kolmogorov-Smirnov test reported a statistic of 0.293 (p-value = 1.33 × 10⁻³⁵), both indicating strong deviations from normality. Given these violations, simple linear regression was deemed unsuitable.

Further, the dataset exhibited significant multicollinearity, as indicated by the Variance Inflation Factor (VIF) values. Several features showed VIFs well above the commonly accepted threshold of 10, such as count\_female (VIF = 55.88), voice\_activity\_female (VIF = 56.20), and count\_male (VIF = 46.79). These values indicate that these variables are highly correlated with other predictors in the model, which can distort the estimation of regression coefficients and reduce model interpretability. Even other features like energy\_male (VIF = 21.14) and voice\_activity\_male (VIF = 42.17) showed substantial collinearity.

To further understand the relationships between features and the target bias score, Pearson's correlation coefficients were calculated. energy\_male showed the strongest positive correlation with the score (r = 0.753), followed by pitch\_male (r = 0.628) and amplitude\_male (r = 0.427). In contrast, features such as energy\_female (r = –0.786) and pitch\_female (r = –0.605) demonstrated strong negative correlations, implying opposing contributions to the bias score. These insights further guided feature selection and model development.

Due to these statistical challenges, namely, violations of normality, homoscedasticity, independence, and multicollinearity, alternative regression approaches were explored. Specifically, symbolic regression and polynomial regression with L2 (Ridge) regularization were selected for constructing the bias detection equation. These methods are more robust against the aforementioned issues and capable of capturing complex, non-linear relationships among variables. The final equation was chosen based on a comparative performance analysis between the two methods, ensuring that the most effective and interpretable model was selected for bias metric estimation.

1. Building and training the equation.

### Building the equation : Method - Symbolic regression.

The training dataset was input into the symbolic regression function for model training, during which various hyperparameter values were systematically evaluated to optimize performance. The parameters tested included population sizes of 500, 1000, and 1500; generation counts of 5, 10, and 15; crossover probabilities (p\_crossover) of 0.6 and 0.7; and point mutation probabilities (p\_point\_mutation) of 0.05 and 0.1. After analyzing the outcomes across these combinations, the optimal configuration was identified as a population size of 500, a point mutation probability of 0.05, a crossover probability of 0.7, and 15 generations. This setup resulted in a Mean Squared Error (MSE) value of 6.64160. The equation generated through symbolic regression under this configuration was subsequently used to compute the Audio Bias Score. The equation produced as follows:

*X0 = number of male audios*

*X1 = number of female audios*

*X2= Voice activity time of male*

*X3 = voice activity of time of female*

*X4 = Standard deviation of Energy levels male*

*X5= Standard deviation of Energy levels female.*

*X6 = Standard deviation of amplitudes male*

*X7 = Standard deviation of amplitudes female*

*X8 = Standard deviation of pitch male*

*X9 = Standard deviation of pitch female*

Following extensive testing, it was observed that when a dataset exhibited gender bias toward male speakers, the Audio Bias Score yielded positive values, while bias toward female speakers resulted in negative values. While this polarity was sufficient to indicate the presence and direction of gender bias, it did not provide a reliable quantification of its magnitude.

To improve interpretability and standardize the metric, the Audio Bias Score was scaled to a range of -10 to 10 using min-max normalization. However, empirical evaluation revealed that the scaled score reliably represented the degree of bias only when gender bias in the dataset was within a 40% threshold. Beyond this point, the score began to exhibit diminishing sensitivity, resulting in a gradual decline in accuracy. This behavior is illustrated in the graphs below:

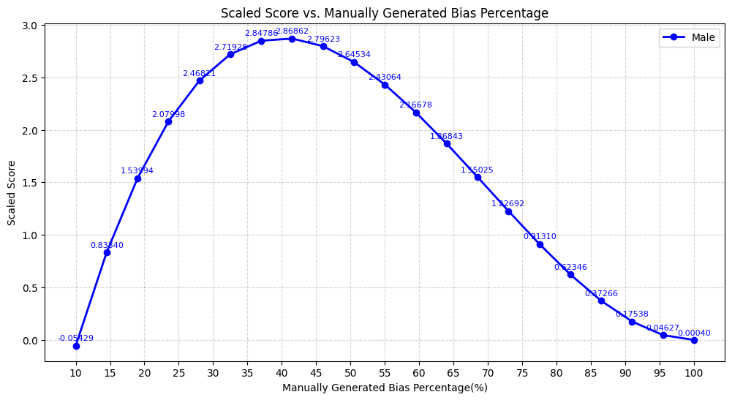


Figure 1 Audio Bias Score computed using the symbolic regression-derived equation versus the percentage of introduced male bias in the dataset.

A graph with red lines

AI-generated content may be incorrect.

Figure 2 Audio Bias Score computed using the symbolic regression-derived equation versus the percentage of introduced female bias in the dataset.

Since the symbolic regression model did not demonstrate the expected performance characteristics, particularly due to its relatively high MSE value, it was concluded that a revised methodology would be necessary to develop a more robust and accurate version of the Audio Bias Score.

### Building the equation : Method - Polynomial Regression with Ridge (L2) regularization.

The augmented dataset was subsequently used to train an Elastic Net Regression model. A grid search over multiple hyperparameter combinations was conducted, where values for alpha were selected from [0.001, 0.01, 0.1, 1.0, 10.0] and l1\_ratio from [0.0, 0.1, 0.5, 0.9, 1.0]. The optimal configuration was found to be alpha = 0.01 and l1\_ratio = 0.0, which corresponds to Ridge regression behavior. This configuration achieved a Mean Squared Error (MSE) of 0.0016 and an R-squared (R²) value of 0.9998, indicating an excellent fit. The resulting regression equation was used to compute the Audio Bias Score. Which is as:

Extensive testing was performed to evaluate the generalization of the derived equation across multiple datasets. As with symbolic regression, the Audio Bias Score exhibited a positive value when the dataset was biased towards male speakers and a negative value when biased towards female speakers, as demonstrated in the graphs provided below.

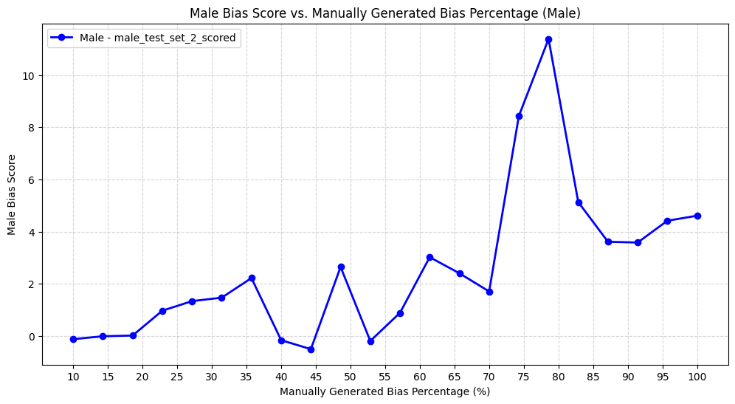


Figure 3 Audio Bias Score computed using the polynomial regression-derived equation versus the percentage of introduced male bias in the dataset.

A graph with a line graph

AI-generated content may be incorrect.

Figure 4 Audio Bias Score computed using the polynomial regression-derived equation versus the percentage of introduced female bias in the dataset.

While both the symbolic and polynomial regression approaches showed this polarity trend, a significant difference emerged in how the scores evolved with increasing levels of bias. Specifically, the score generated by the polynomial regression model increased consistently and linearly as the level of gender bias intensified. In contrast, the symbolic regression-based score exhibited a decline beyond a 40% bias threshold, thereby compromising its interpretability in high-bias scenarios.

Despite the improved consistency in score trends, the raw output values from the polynomial regression model were not directly interpretable in terms of quantifying the degree of bias. To resolve this, min-max scaling was applied to constrain the Audio Bias Score within a normalized range, making the output more interpretable.

A polarity-aware scaling method was introduced to enhance this normalization. When the raw Audio Bias Score was positive, the maximum bound was computed by minimizing female-associated feature values, while the minimum bound was determined using equal values for both male and female features. Conversely, if the score was negative, the process was reversed. This polarity-sensitive normalization approach ensured that the scaled Audio Bias Score exhibited greater consistency and interpretability across varying degrees of dataset bias.

Upon comparative evaluation, the equation derived through polynomial regression with Ridge regularization was found to be superior to the symbolic regression-based model. It provided a more accurate, consistent, and interpretable measure of gender bias in audio datasets, thereby establishing it as the preferred formulation for the Audio Bias Score metric.

TESTING.

To evaluate the generalization capability and robustness of the proposed Audio Bias Score, the final regression equation was tested on 10 diverse datasets spanning multiple languages and recording conditions. These datasets were manually manipulated to reflect varying levels of gender bias, after which the Audio Bias Score was computed and scaled to a standardized range between -10 to 10 using the polarity-aware min-max normalization approach described previously.

Following the score computation and scaling, several statistical and regression-based evaluation metrics were calculated to assess the accuracy, consistency, and correlation strength of the Audio Bias Score. These metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Normalized Mean Squared Error (NMSE), R-squared (R²), as well as Pearson’s Correlation Coefficient, Spearman’s Rank Correlation, and Kendall’s Tau Rank Correlation. The results are presented in Table I and Table II.

TABLE I : Performance metrics of the Audio Bias Score equation across multiple datasets. Abbreviations: LBS – LibriSpeech, MLT – Multilingual LibriSpeech, CMV – Common Voice

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset** | **MSE** | **RMSE** | **MAE** | **NMSE** |  |
| LBS | 0.07173 | 0.26784 | 0.1724 | 0.00209 | 0.9979 |
| MLT: Italian | 0.02688 | 0.16395 | 0.1149 | 0.00078 | 0.9992 |
| MLT: Portuguese | 1.23754 | 1.11245 | 0.4622 | 0.03622 | 0.9638 |
| MLT: Polish | 0.35646 | 0.59704 | 0.3112 | 0.01043 | 0.9896 |
| CMV : Hakha Chin | 0.57049 | 0.75530 | 0.5480 | 0.01669 | 0.9833 |
| CMV : Chuvash | 1.01335 | 1.00665 | 0.7063 | 0.02965 | 0.9703 |
| CMV : Welsh | 0.00097 | 0.03125 | 0.0254 | 2.85968 | 0.9998 |
| CMV : Kurmanji | 0.01545 | 0.12433 | 0.0990 | 0.00045 | 0.9995 |
| TedLium | 0.26426 | 0.51406 | 0.3697 | 0.00773 | 0.9923 |
| The AMI Corpus | 0.00503 | 0.07096 | 0.0517 | 0.00014 | 0.99985 |

TABLE II : Performance metrics of the Audio Bias Score equation across multiple datasets cont... Abbreviations: LBS – LibriSpeech, MLT – Multilingual LibriSpeech, CMV – Common Voice.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Pearson’s Correlation** | **Spearman’s Rank**  **Correlation** | **Kendall Tau**  **Rank Correlation** |
| LBS | 0.9991 | 0.99951 | 0.99534 |
| MLT: Italian | 0.9997 | 0.99983 | 0.99767 |
| MLT: Portuguese | 0.9848 | 0.98654 | 0.95348 |
| MLT: Polish | 0.9954 | 0.99485 | 0.97558 |
| CMV : Hakha Chin | 0.9972 | 0.99752 | 0.98255 |
| CMV : Chuvash | 0.9962 | 0.99760 | 0.98486 |
| CMV : Welsh | 0.9993 | 0.999999 | 0.999999 |
| CMV : Kurmanji | 0.9999 | 0.99999 | 0.99999 |
| TedLium | 0.9970 | 0.99756 | 0.98604 |
| The AMI Corpus | 0.9999 | 0.99999 | 0.99999 |

The results demonstrate the effectiveness and reliability of the Audio Bias Score in capturing and quantifying gender bias across diverse datasets. The high R² values (ranging from 0.9638 to 0.99985) indicate that the metric closely approximates the actual degree of bias. Moreover, the strong correlation values, particularly for the Pearson, Spearman, and Kendall Tau coefficients highlight the Audio Bias Score's consistency in ranking and aligning with the actual bias configurations.

VALIDATION.

To validate the effectiveness and reliability of the proposed Audio Bias Score, the Word Error Rate (WER) was employed as a comparative benchmark. WER is a widely accepted metric in evaluating the performance and bias of speech recognition systems and has been commonly used to measure disparities in recognition accuracy between male and female speakers. It reflects the combined influence of the dataset and the underlying speech recognition model on gender-related performance gaps.

In this study, WER was calculated using the Whisper-tiny speech recognition model developed by OpenAI [33], due to its multilingual support and general robustness. Each dataset split was transcribed using this model, and WER values were computed separately for male and female speakers.

The Audio Bias Score, in contrast to WER, is designed to reflect the inherent bias in the dataset itself rather than the bias introduced by the model. The Audio Bias Score ranges from –10 to +10, where positive values indicate bias towards male speakers, and negative values indicate bias towards female speakers. This score was computed based on features extracted from each dataset split using the same pipeline.

TABLE III : side-by-side comparison between the WER values for male and female speakers Vs the dataset bias direction inferred from the Audio Bias Score. Abbreviations: LBS – LibriSpeech, MLT – Multilingual LibriSpeech, CMV – Common Voice.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Datasets** | **WER male** | **WER female** | **Bias defined by WER** | **Bias score** | **Definition of the score** |
| LBS: Dev split | 20.05 | 26.3 | System : Male biased | 0.650504 | Dataset : Male Biased |
| LBS: Test-other split | 16.0625 | 19.0588 | System : Male biased | 1.263191 | Dataset : Male Biased |
| LBS: Test-clean split | 22.45 | 24.75 | System : Male biased | 0.811159 | Dataset : Male Biased |
| LBS Train split | 31.5104 | 32.0797 | System : Male biased | 0.713163 | Dataset : Male Biased |
| MLT : Portuguese | 1.148 | 1.2917 | System : Male biased | -4.79006 | Dataset : Female Biased |
| MLT : Polish | 1.222 | 1.2009 | System : Female biased | 6.279849 | Dataset : Male Biased |
| CMV : Hakha Chin Train split | 1.17 | 1.2114 | System : Male biased | 1.259123 | Dataset : Male Biased |
| CMV : Hakha Chin other-split | 1.1913 | 1.1497 | System : Female biased | -5.24067 | Dataset : Female Biased |
| CMV : Hakha Chin validated split | 1.1816 | 1.1599 | System : Female biased | -0.23507 | Dataset : Female Biased |
| CMV : Chuvash test split | 1.2571 | 1.2261 | System : Female biased | -3.51048 | Dataset : Female Biased |
| CMV : Chuvash validated split | 1.2433 | 1.2966 | System : Male biased | 2.512189 | Dataset : Male Biased |
| CMV : Sorani validated split | 1.2121 | 1.4281 | System : Male biased | 8.25787 | Dataset : Male Biased |
| CMV : Western Frisian other split | 1.6055 | 1.3719 | System : Female biased | -4.88802 | Dataset : Female Biased |
| CMV : Western Frisian validated split | 2.7478 | 1.4275 | System : Female biased | -6.17234 | Dataset : Female Biased |
| CMV : Indonesian validated split | 1.5453 | 1.479 | System : Female biased | -7.78833 | Dataset : Female Biased |
| CMV : Latgalian validated split | 1.4751 | 1.4529 | System : Female biased | -6.42433 | Dataset : Female Biased |

In most cases, the bias direction identified by the WER values and the Audio Bias Score are aligned, supporting the validity of the Audio Bias Score as a metric to identify gender bias in audio datasets. However, a few discrepancies were observed, where the system-level bias detected via WER did not match the dataset-level bias detected via the Audio Bias Score. This highlights an important distinction: WER captures the compounded effects of both dataset and model bias, whereas the Audio Bias Score is specifically engineered to isolate and quantify the dataset bias. Thus, this comparison emphasizes the potential of the Audio Bias Score to serve as a reliable and model-independent measure for identifying gender bias in audio datasets.

DISCUSSION.

While the proposed Audio Bias Score provides a novel approach for quantifying gender bias in audio datasets, there are several limitations that warrant further investigation. Firstly, the current metric is limited to evaluating gender bias and does not account for other demographic attributes such as age or race. These factors can also contribute to disparities in audio-based systems and should be incorporated in future iterations of the metric.

Secondly, the effectiveness of the proposed metric relies on the availability of speaker metadata, particularly gender labels. However, many public audio datasets do not provide this information, which limits the metric’s applicability. Additionally, although most audio-based systems are typically trained on single-speaker utterances, often obtained by applying speaker diarization to multi-speaker recordings, this study focuses only on pre-segmented, single-speaker data. As such, it does not assess the metric’s performance in the context of multi-speaker audio. Future research should explore how the metric can be extended or adapted for use with multi-speaker datasets.

In conclusion, this study takes a critical step toward evaluating inherent gender bias in audio datasets by introducing a dedicated metric that operates independently of downstream model performance. As AI systems continue to influence high-stakes decision-making, ensuring fairness at the dataset level is essential. This research contributes to the broader effort of building transparent and equitable AI systems by highlighting the need to identify and mitigate bias at the source, within the data itself.

CONCLUSION.

This study introduces the Audio Bias Score, a metric designed to quantify gender bias in audio datasets. By leveraging raw audio features and avoiding dependency on downstream models or algorithms, the metric offers a dataset-focused perspective on bias detection. Among several modeling approaches tested, polynomial regression with L2 regularization (Ridge regression) produced the most consistent and interpretable results.

Validation of the metric against the established Word Error Rate (WER), calculated using OpenAI’s Whisper-tiny speech recognition model, demonstrated a strong alignment between the Audio Bias Score and model-based bias indicators. Importantly, the proposed metric isolates dataset-induced bias, offering a critical advantage over traditional system-level metrics like WER.

The Audio Bias Score serves as a foundational step toward the development of fair and inclusive audio-based AI systems. By enabling developers and researchers to identify and quantify gender bias at the dataset level, it paves the way for the creation of more equitable machine learning models and tools.

REFERENCES.

|  |  |
| --- | --- |
| [1] | A. Alrosan, M. Abdel- Atv, M. Hafez, S. Alkhazaleh, M. A. Deif and R. ELGohary, "Parkinson's Disease Detection Based on Vocal Biomarkers and Machine Learning Approach," in *International Telecommunications Conference (ITC-Egypt)*, Cairo, Egypt, 2024. |
| [2] | H. W. Ko and S. Kwon, "Optimization of voice biomarkers to predict Alzheimer’s disease," *Alzheimer's & Dementia,* vol. 19, no. 15, p. 79907, 2023. |
| [3] | S. Chen, L. Li, S. Han, W. Luo, W. Wang, Y. Yang, X. Wang, W. Zhang, M. Chen and Z. Wang, "Review of voice biomarkers in the screening of neurodegenerative diseases," *Interdisciplinary Nursing Research,* vol. 3, pp. 190-198, 2024. |
| [4] | V. Urovi, H. Strik, H. van Helvoort, S. O. Simons and I. Michels, "Vocal biomarkers in COPD: capturing disease severity using voice," *European Respiratory Journal,* vol. 60, p. 1591, 2022. |
| [5] | T. Marinopoulou, A. Lalas, K. Votis and D. Tzovaras, "An AI-powered Acoustic Detection System Based on YAMNet for UAVs in Search and Rescue Operations," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings, InterNoise23*, Chiba, Japan, 2023. |
| [6] | R. Jahangir, Y. Wah Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Expert Systems with Applications,* vol. 171, p. 114591, 2021. |
| [7] | M. Best and A. Rao, "Understanding algorithmic bias and how to build trust in AI," 18 January 2022. [Online]. Available: https://www.pwc.com/us/en/tech-effect/ai-analytics/algorithmic-bias-and-trust-in-ai.html. |
| [8] | M. Best, "AI bias is personal for me. It should be for you, too," 30 June 2021. [Online]. Available: https://www.pwc.com/us/en/tech-effect/ai-analytics/artificial-intelligence-bias.html. |
| [9] | A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences,* vol. 117, no. 14, pp. 7684-7689, 2020. |
| [10] | G. Attanasio, B. Savoldi, D. Fucci and D. Hovy, "Twists, Humps, and Pebbles: Multilingual Speech Recognition Models Exhibit Gender Performance Gaps," in *2024*, Miami, Florida, USA, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. |
| [11] | C. S. Greenberg, L. P. Mason, S. O. Sadjadi and D. A. Reynolds, "Two decades of speaker recognition evaluation at the national institute of standards and technology," *Computer Speech & Language,,* vol. 60, p. 101032, 2020. |
| [12] | W. Hutiri, T. Patel, A. Ding and O. Scharenborg, "As Biased as You Measure: Methodological Pitfalls of Bias Evaluations in Speaker Verification Research.," in *Proceedings of Interspeech 2024*, Kos, Greece, 2024. |
| [13] | J. Meyer, L. Rauchenstein, J. D. Eisenberg and N. Howell, "Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications," in *Proceedings of the Twelfth Language Resources and Evaluation Conference* , Marseille, France, 2020. |
| [14] | s. Feng , O. Kudina, B. M. Halpern and O. Scharenborg, "Quantifying Bias in Automatic Speech Recognition," *arXiv preprint arXiv:2103.15122,* 2021. |
| [15] | M. Z. Boito, L. Besacier, N. Tomashenko and Y. Est`eve, "A Study of Gender Impact in Self-supervised Models for Speech-to-Text Systems," in *Proceedings of Interspeech 2022*, Incheon, Korea., 2022. |
| [16] | F. Burkhardt, B. T. Atmaja, A. Derington, F. Eyben and B. Schuller, "Check Your Audio Data: Nkululeko for Bias Detection," in *27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, Hsinchu City, Taiwan, 2024. |
| [17] | M. Shabbir, A. Hussain and M. M. Khan, "Age and Gender Estimation Through Speech: A Comparison of Various Techniques," in *2023 18th International Conference on Emerging Technologies (ICET)*, Peshawar, Pakistan, 2023. |
| [18] | A. Ghosal, C. Pathak, P. Singh and S. Dutta, "Voice-Based Gender Identification Using Co-occurrence-Based Features," *Computational Intelligence in Pattern Recognition,* pp. 947-956, 2020. |
| [19] | E. Priya, S. J. Priyadharshini, P. S. Reshma and S. Sashaank , "Temporal and spectral features based gender recognition from audio signals," in *2022 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, Chennai, India, 2022. |
| [20] | Y. Ali, E. Noorsal, N. F. Mokhtar, S. Z. Md Saad, M. H. Abdullah and L. C. Chin, "Speech-based gender recognition using linear prediction and mel-frequency cepstral coefficients," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 28, no. 2, p. 753, 2022. |
| [21] | S. Fahmeeda, M. A. Ayan, M. Shamsuddin and A. Amreen, "Voice Based Gender Recognition Using Deep Learning," *International Journal of Innovative Research & Growth,* vol. 3, pp. 649-654, 2022. |
| [22] | S. Srivastava, H. Sharma and D. Garg, "Comparative Study of Machine Learning Algorithms for Voice based Gender Identification," in *2022 International Conference on Edge Computing and Applications (ICECAA)*, Tamilnadu, India, 2022. |
| [23] | Z. Zhang , R. Li and K. Chen, "Speaker Gender Recognition Based on Semi-Supervised Learning," in *2024 3rd International Conference on Computing, Communication, Perception and Quantum Technology (CCPQT)*, Zhuhai, China, 2024. |
| [24] | I. A. Destina and E. Hartati, "An Analysis of intonation pattern in the pre-service english teacher talks," *FRASA : English Education and Literature Journal,* vol. 3, no. 2, pp. 64-71, 2022. |
| [25] | S. Ekeruke, "Stress Patterns and Intonations among the Annang Language Speaking Students of Faculty of Arts, Akwa Ibom State University," *Journal of Cmmunication and Culture,* vol. 12, no. 3, pp. 163-172, 2024. |
| [26] | B. Ludusan, M. Heldner and M. Wlodarczak, "Exploring the role of formant frequencies in the classification of phonation type," in *Proceedings of 20th International Congress of Phonetic Sciences*, Prague, Czech Republic, 2023. |
| [27] | A. Bailey and M. D. Plumbley, "Gender Bias in Depression Detection Using Audio Features," in *Proceedings of the 29th European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, 2021. |
| [28] | R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers and G. Weber, "Common Voice: A Massively-Multilingual Speech Corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, 2020. |
| [29] | V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, 2015. |
| [30] | V. Pratap , Q. Xu, A. Sriram, G. Synnaeve and R. Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Interspeech 2020*, Shanghai, China, 2020. |
| [31] | A. Rousseau, P. Deléglise and Y. Estève, "TED-LIUM: an Automatic Speech Recognition dedicated corpus," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012. |
| [32] | I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, . T. Hain, J. Kadlec, . V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma and . P. Wellner, "The AMI meeting corpus," in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008. |
| [33] | A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML'23: Proceedings of the 40th International Conference on Machine Learning*, Honolulu Hawaii USA, 2023. |