IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 February 2026, accepted 6 March 2026, date of publication 17 March 2026, date of current version 20 March 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3674991

RESEARCH ARTICLE

EuqAud: Detecting Gender Bias in Audio Datasets Using Polynomial Regression-Based Metric

SRIWANTHI JAYAWARDENA, PRASANNA S. HADDELA, THISARA SHYAMALEE, AMANDI EKANAYAKE, THARUSHI MUDALIGE, AND IMESHA DHANAWARDHANA

Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe 10115, Sri Lanka

Corresponding author: Sriwanthi Jayawardena (sriwanthij@gmail.com)

ABSTRACT With the growing adoption of audio based AI systems in high-stakes domains such as healthcare, law enforcement, and social media, ensuring fairness particularly regarding gender bias has become critically important. While prior work on fairness has predominantly focused on disparities in model performance, bias inherent in training datasets remains underexplored. To address this gap, we propose EuqAud, a novel, pre-trained and traceable fairness metric that quantifies gender bias in audio datasets using raw acoustic features such as pitch, energy, amplitude, and voice activity. Unlike methods dependent on demographic labels such as race, age or language, EuqAud is designed to be demographic and language agnostic, enhancing its applicability across diverse contexts. The score is computed using an equation derived from polynomial regression with L2 regularization (Ridge regression), yielding robust and generalizable outputs. It spans a range from -10 to 10, where 0 denotes neutral, positive scores indicate male dominant bias, and negative scores reflect female dominant bias. For clarity, bias severity is categorized into three tiers: Neutral (EuqAud < 2), Moderate Bias ($2 \leq$ EuqAud ≤ 6), and Strong Bias (EuqAud > 6). Evaluation across multiple datasets demonstrates high predictive performance, with R^2 values between 0.95 and 0.99. By focusing on dataset level bias rather than model outcomes, EuqAud offers a scalable and rigorous solution for advancing fairness in audio-based AI systems.

INDEX TERMS Audio datasets, bias detection, EuqAud, gender bias, polynomial regression, responsible AI.

I. 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, with audio being one of the most prominent. 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 Dablain et al. [7], bias arises from unequal representation or distortion of protected groups within a dataset, leading to disparities in how models learn and generalize across those groups. Such bias can arise from unrepresentative training data that embeds existing social prejudices, flawed algorithmic design, or real-world interactions that reinforce disparities [8].

Recent advances in dataset-level fairness research emphasize that harmful model behavior often originates from structural imbalances within the data itself. Prior works such as Gender Shades Buolamwini and Gebru [9], Suresh and Guttag [10], Raji et al. [11], and Dablain et al. [7] demonstrate that dataset composition can directly propagate algorithmic harm, motivating the need for pre-model, dataset-level bias evaluation. Additional works in dataset governance, such as Datasheets for Datasets [12] and Data Cards [13],

The associate editor coordinating the review of this manuscript and approving it for publication was Sajid Ali.

further underscore the importance of systematic dataset auditing prior to model development.

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 [14].

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 [15].

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 underrepresented training datasets. The primary objective is to develop a metric that can quantify the extent of gender bias in audio datasets, independently 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.

To address this challenge, this study proposes the EuqAud metric and evaluates its effectiveness in detecting dataset-level gender bias in audio data. The methodology,evaluation and limitations of the proposed approach are presented throughout the remainder of this paper. The remainder of this paper is organized as follows, starting under Section II, relevant related works and methodologies for bias detection in audio-based AI systems are discussed. Section III details the development of the proposed EuqAud metric, including feature selection, dataset construction, and regression modeling. Section IV describes the testing and validation procedures used to evaluate the accuracy and generalization capability of EuqAud where the equation is tested on real world data. Section V discusses the findings, limitations, and implications for future research. Finally, Section VI concludes the study and outlines potential directions for extending EuqAud to other demographic attributes and multimodal data.

II. 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) [16]. 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 such as EER, FPR, FNR [17].

Specifically these methods each capture the following:

- G2mindiff captures the gap between the base performance of a demographic group and the best-performing group.
- G2avg ratio compares a group's performance to the average performance across all groups.
- G2avg log ratio offers a scaled log-based perspective on these performance gaps.

$$G2mindiff = b_g - b_m \tag{1}$$

$$G2avg\ Ratio\ (b_g) = \frac{b_g}{\bar{b}} \tag{2}$$

$$G2avg\ log\ Ratio\ (b_g) = -ln\left(\frac{b_g}{\bar{b}}\right) \tag{3}$$

→ b_g is the base metric for the demographic group being evaluated
→ b_m is the metric of the best performing group
→ $\bar{b}$ is the average performance across all demographic groups

For speech recognition systems, metrics such as Character Error Rate (CER) [18] and Word Error Rate (WER) [19], [20] are commonly used. CER measures the difference between predicted and actual characters in transcriptions, while WER compares predicted and actual word level transcriptions.

$$CER = \frac{Number\ of\ in\ correctly\ predicted\ characters}{Total\ number\ of\ characters\ in\ the\ transcription} \tag{4}$$

$$WER = \frac{Number\ of\ in\ correctly\ predicted\ words}{Total\ number\ of\ words\ in\ the\ transcription} \tag{5}$$

Whereas 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 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. introduced the Nkululeko framework [21], 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 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.

whereas most of the 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, traceable, and gender specific bias score grounded entirely in the structure of the audio data itself.

III. BUILDING EuqAud

The section presents the systematic development of EuqAud. The methodology comprises a series of structured steps, each represented as a dedicated subsection in this section. These steps include what the metric is intended to measure, the selection of relevant features informed by prior research, construction of a training dataset through controlled augmentation, and regression based modeling to derive the final bias equation. Emphasis is placed on ensuring robustness, interpretability, and independence from downstream model performance. Each component is discussed in detail to highlight the design rationale and demonstrate the generalizability of EuqAud across diverse audio corpora.

A. WHAT DOES EuqAud MEASURE

The study defines bias as the systematic underrepresentation of one gender's acoustic features relative to the other, following the framing of what bias is in Dablain et al. [7]. This definition captures dataset-level phenomena such as reduced speaking intensity or vocal attenuation, which can make acoustic presence of one gender less prominent even when sample counts appear balanced.

EuqAud is designed to directly measure this acoustic underrepresentation, allowing developers to identify when the underlying data may lead to unequal downstream model performance across genders. Importantly, EuqAud is computed entirely from dataset-level acoustic statistics; no ASR system, classifier, or model predictions are used or required at any stage of metric construction.

Once trained, EuqAud becomes a fixed, closed-form regression equation. It does not undergo retraining for new datasets. Instead, the same final equation is applied directly to any audio dataset that includes reliable metadata linking each audio clip to a speaker's gender. Also as the metric relies only on fundamental acoustic statistics and not on language-specific features, EuqAud can be applied consistently across datasets from different languages and recording environments.

B. SELECTION OF FEATURES FOR EQUATION BUILDING

The selection of features for building EuqAud 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 [22], amplitude [23], energy [24], formants, intonations, and Mel-Frequency Cepstral Coefficients (MFCCs) [25], [26], [27], [28].

However, intonations [29], [30] and formants [31] are dependent on the language and can vary among cultures and ethnic groups. Since EuqAud 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., models using raw audio are more robust to gender bias than those based on hand-crafted features, such as mel-spectrograms [32]. Therefore, EuqAud 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, EuqAud 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.

C. EXTRACTING, BUILDING AND PROCESSING THE DATASETS

To build EuqAud equation, a base dataset was first created by extracting gender-specific acoustic statistics from multiple real-world speech corpora. Each row in the base dataset corresponds to one dataset split combination and contains aggregated male and female counts, voice activity duration, and the standard deviations of pitch, energy, and amplitude. This base dataset represents the natural, unmodified acoustic distributions present in real-world data. The datasets used to build the base dataset included Common Voice [33], LibriSpeech [34], LibriSpeech Multi-lingual [35], TED-LIUM [36], and the AMI Meeting Corpus [37]. 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 acoustics features used in the study, defining a ground truth bias value required the creation of controlled conditions that systematically manipulate gender representation.

To achieve this, a feature level augmentation technique was applied that preserved the number of samples while modifying acoustic prominence of each gender.

1. For the first half of the dataset, the extracted values for male features: Voice activity, pitch, energy and amplitude were kept constant, while the corresponding female acoustic features were gradually reduced in increments of 5% by multiplying original values by 0.95, 0.90, 0.85, …for successive steps, while maintaining a constant number of samples.
2. For the second half, the values for female acoustic features remained unchanged, while the male acoustic features were gradually reduced in increments of 5% using the same multiplicative procedure under the same conditions of the constant sample size.

The 5% incremental feature reduction simulates real-world scenarios where one group's vocal characteristics appear less prominently in the training distribution, independent of sample count.

This increment used in the study was selected as a small, controlled proportional perturbation designed to introduce gradual and interpretable variations in the signal while preserving natural acoustic characteristics through the analysis of prior work.

However, prior work does not prescribe a specific 5% step size, the speech-processing literature commonly employs small multiplicative perturbations of similar magnitude. For example, pitch- and frequency-based augmentations often use modest scaling factors such as ± 3% to 6% to generate realistic variants for robust training [38], [39]. Likewise, amplitude and gain perturbations in environmental-robustness studies frequently operate within ±1% to 10% to simulate changes in microphone conditions and loudness without producing unnatural artefacts [40], [41]. Broader augmentation methods such as time-scaling or frequency warping typically use moderate ranges like 0.9 to 1.1 (±10%) or even 0.8 to 1.2 [42], [43], demonstrating that proportional manipulations are a well-established strategy for generating realistic variability in speech signals. Following this principle, the 5% step size provides a fine-grained and conservative perturbation that enables progressive bias analysis while remaining consistent and in range with the magnitudes used in prior augmentation and robustness studies.

By keeping the number of male and female samples constant, this approach isolates bias arising from distorted feature distributions rather than class imbalance. In this setup, bias arises purely from acoustic underrepresentation, aligning with how demographic underrepresentation manifests in practical audio corpora.

The resulting bias scores correspond to the degree of reduction applied(0.0, 0.5, 1.0,…) indicating the level of bias in the dataset that EuqAud will quantify. Each of these assigned bias scores serves as the target variable for training the regression model. During training, the polynomial Ridge regression model used the distorted acoustic features as the input values and the numeric bias score as the ground truth target. This enables EuqAud to predict the bias score for any given datasets which reflects the degree and the direction of gender bias.

Although this controlled feature-level augmentation enables the creation of datasets with known bias levels, it may not fully capture the complex patterns through which demographic imbalance emerges in natural datasets. Real-world bias can arise from multiple interacting factors such as recording conditions, sociolinguistic differences, demographic participation patterns, and data collection pipelines. Therefore, the synthetic perturbations used in this study should be interpreted as controlled approximations that allow the regression model to learn a structured mapping between acoustic disparities and bias scores. To mitigate this limitation, EuqAud was subsequently evaluated on multiple real-world datasets with naturally occurring gender distributions, demonstrating that the learned equation generalizes beyond the synthetic training conditions.

D. SELECTION OF MODELING APPROACH FOR EuqAud DEVELOPMENT

The training dataset was generated by applying a controlled data augmentation technique to the base dataset which contains aggregated acoustic statistics such as counts, voice activity, and standard deviations of pitch, energy, and amplitude extracted from real-world speech corpora. This augmentation procedure systematically reduced either male or female acoustic feature values in fixed increments of 5% while keeping sample counts constant, generating a series

of datasets with known, controlled levels of representational bias.

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 1.027, 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 $= 1.95 \times 10^{-13}$) and an F-statistic of 9.741 (p-value $= 8.54 \times 10^{-15}$). 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 $= 2.41 \times 10^{-10}$), and the Kolmogorov-Smirnov test reported a statistic of 0.293 (p-value $= 1.33 \times 10^{-35}$), 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, such as violations of normality, homoscedasticity, independence, and multicollinearity, alternative regression approaches were explored, after which Polynomial regression with L2 (Ridge) regularization was selected for constructing the bias detection equation.

E. BUILDING THE EQUATION USING POLYNOMIAL REGRESSION WITH RIDGE(L2) REGULARIZATION

The augmented dataset was used to train an Elastic Net Regression model. The model was trained to predict ground truth bias scores ranging from 0 to 10 based on input acoustic feature statistics. Producing a continuous EuqAud value that quantifies dataset level gender bias.

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 (R^2) value of 0.9998, indicating an excellent fit. The resulting regression equation was used as EuqAud.

The resulting equation which is EuqAud presents as follows:

$$\begin{aligned} EuqAud = {} & -0.0125 + (0.0001C_{male}) + \left(-0.0001C_{female}\right) \\ & + (0.0005V_{male}) + \left(-0.0005V_{female}\right) \\ & + (0.0044E_{male}) + \left(-0.0034E_{female}\right) \\ & + (-0.0004A_{male}) + \left(-0.0004A_{female}\right) \\ & + (-0.0334P_{male}) + \left(-0.0325P_{female}\right) \\ & + \left(0.0002P_{male}^2\right) + \left(-0.0002P_{male}P_{female}\right) \\ & + (0.0002P_{female}^2) \end{aligned}$$

$$\begin{aligned} C_{male} &: Count\ of\ male\ audios \\ C_{female} &: Count\ of\ female\ audios \\ V_{male} &: Voice\ activity\ male \\ V_{female} &: Voice\ activity\ female \\ E_{male} &: Standard\ deviation\ of\ energy\ male \\ E_{female} &: Standard\ deviation\ of\ energy\ female \\ A_{male} &: Standard\ deviation\ of\ Amplitude\ male \\ A_{female} &: Standard\ deviation\ of\ Amplitude\ female \\ P_{male} &: Standard\ deviation\ of\ Pitch\ male \\ P_{female} &: Standard\ deviation\ of\ Pitch\ female \end{aligned} \quad (6)$$

The equation was extensively tested to evaluate the generalization of EuqAud across multiple synthetic as well as real-world datasets. Throughout the testing the regression outputs consistently reflected the intended direction of bias, producing positive EuqAud values when male acoustic features dominated and a negative value when female acoustic features dominated, as demonstrated in Figure 1 and Figure 2. EuqAud exhibited a positive value when the datasets were biased toward male speakers and negative values when biased towards female speakers.

EuqAud increased consistently as the level of gender bias intensified in either direction. This behavior demonstrated the model's ability to capture proportional relationships between input features and the resulting bias score.

However, the raw output values of EuqAud were not directly interpretable in terms of quantifying the severity of bias. To address this, min–max scaling technique was applied to normalize EuqAud to the range [−10, 10], enhancing interpretability and consistency across datasets.

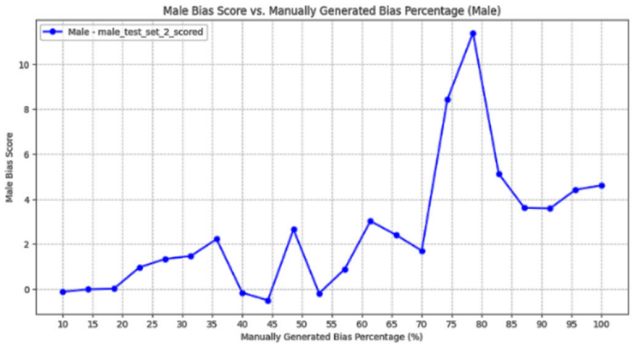

FIGURE 1. EuqAud score as a function of the manually specified male bias percentage in the dataset.

The final normalized EuqAud score spans from −10 to 10, where 0 indicates a neutral dataset, positive values indicate male bias, and negative values indicate female bias. To apply this normalization, different bounds were used based on the sign of the raw score. When the raw EuqAud value was positive (male biased), the maximum bound was set by minimizing female associated feature values, and the minimum bound was set using equal values for both male and female features. Conversely, when the raw score was negative (female biased), the process was reversed. This min-max scaling strategy ensured that the score reflected both the direction and severity of bias consistently across datasets, making EuqAud intuitive for comparative analysis and practical auditing.

To further aid interpretability, a tier-based classification system was developed based on the magnitude of the scaled EuqAud score: Strong Bias (>6), Moderate Bias (2–6), and Neutral (<2). The sign of the score indicates the direction of bias, positive for male and negative for female, providing a clear and traceable assessment of dataset fairness.

It is important to emphasize that EuqAud is a fixed regression equation. Once trained, the model is not retrained for each new dataset; instead, the same polynomial Ridge regression equation is applied directly to any dataset's extracted acoustic features to compute its EuqAud score.

F. EuqAud BIAS MEASUREMENT ALGORITHM

EuqAud follows a stepwise process to calculate bias. Therefore, to improve reproducibility and clarity of implementation, this section provides Algorithm 1 which presents the step-by-step procedure used to compute the EuqAud bias score.

IV. EVALUATION OF EuqAud

The evaluation of EuqAud is conducted through a multi-stage process designed to assess its mathematical accuracy, sensitivity to real-world dataset imbalance, interpretability through its tiered bias structure, and relevance to downstream model behavior and system-level WER disparities. The final stage compares EuqAud against human perceptual judgements to ensure that the fairness metric remains traceable and intuitively meaningful.

Algorithm 1 Calculating EuqAud

Require: $C_{male}, C_{female}, V_{male}, V_{female}, E_{male}, E_{female}, A_{male}, A_{female}, P_{male}, P_{female}$

1. $EuqAud_{raw} \leftarrow$ *Computed using Equation* 6.
2. $EuqAud \leftarrow \text{min\_max\_scale}(EuqAud_{raw}, -10, 10)$
3. **If** $|\ EuqAud\ | < 2$ **then**
   Bias ← Neutral
4. **else if** $2 \leq |\ EuqAud\ | \leq 6$ **then**
   Bias ← Moderate
5. **else**
   Bias ← Strong
6. **end if**
7. **return** *EuqAud*

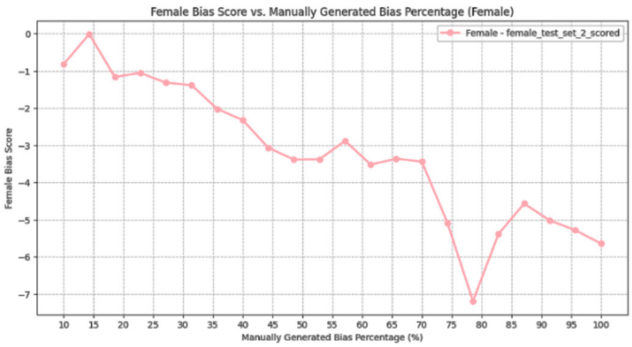

FIGURE 2. EuqAud score as a function of the manually specified female bias percentage in the dataset.

Importantly, because no established ground-truth method exists for measuring dataset-level gender bias in audio corpora, the validation strategy necessarily relies on multiple complementary proxies, including performance disparities (F1-score), representation patterns (gender/VAD dominance), synthetic manipulations, and WER behavior. Every stage captures different aspects of bias which when combined provide a near comprehensive validation framework. This multi-angle evaluation is essential for demonstrating that EuqAud reliably captures dataset bias in a model-agnostic and empirically verifiable manner.

A. REGRESSION-LEVEL EVALUATION OF EuqAud

To evaluate the generalization capability and robustness of the proposed EuqAud, it was tested on 10 diverse datasets spanning multiple languages and recording conditions.

A key challenge in this process was the lack of absolute ground truth labels for bias. To address this, the datasets were manually manipulated to reflect varying levels of gender bias, allowing the assignment of known ground truth labels.

The augmentation was designed to ensure that the testing datasets remained independent from the training datasets,

which had been previously augmented in 5% increments. To achieve this, the test datasets were augmented using varying levels of reduction ranging from 1% to 10%, applied to the extracted acoustic features rather than sample counts. Specifically, for each test dataset, one gender's features (voice activity, pitch, energy, amplitude) were gradually reduced while the other gender's features remained unchanged. For example, LibriSpeech subsets were manipulated by 2% per step, while the TedLium dataset was augmented by 7%, with corresponding ground truth scores assigned based on the level of feature reduction (e.g., LibriSpeech scores: 0.2, 0.4…; TedLium scores: 0.7, 1.4…). This approach preserved natural sample distributions while introducing controlled variations in acoustic prominence, allowing EuqAud to be tested on realistic yet systematically biased conditions

Following the augmentation process, EuqAud was calculated for each feature set. Following the score computation, several statistical and regression-based evaluation metrics were calculated to assess the accuracy, consistency, and correlation strength of EuqAud. These metrics include R-squared (R^2), Pearson's Correlation Coefficient, and Spearman's Rank Correlation, the results are presented in Table 1.

TABLE 1. Regression performance and correlation analysis of the EuqAud equation across multiple datasets. Abbreviations: LBS – LibriSpeech, MLT – Multilingual LibriSpeech, CMV – Common Voice.

Dataset	r^2	Pearson's Correlation	Spearman's Rank Correlation
LBS	0.998	0.999	0.999
MLT: Italian	0.999	0.999	0.999
MLT: Portuguese	0.964	0.984	0.986
MLT: Polish	0.990	0.995	0.994
CMV : Hakha Chin	0.983	0.997	0.997
CMV : Chuvash	0.970	0.996	0.997
CMV : Welsh	1.000	0.999	0.999
CMV : Kurmanji	0.015	0.999	0.999
TedLium	0.264	0.997	0.997
The AMI Corpus	0.005	0.999	0.999

The results demonstrate the effectiveness and reliability of EuqAud in capturing and quantifying gender bias across diverse datasets. The high R^2 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 Pearson, and Spearman highlight EuqAud's consistency in ranking and aligning with the actual bias configurations.

B. COMPARISON BETWEEN SIMPLE BASELINE METRICS AND EuqAud

Baseline metrics provide a useful point of reference by demonstrating how well a proposed method performs relative to straightforward statistical measures derived directly from the dataset. By comparing EuqAud with baseline approaches, it becomes possible to assess whether the proposed metric captures bias patterns more comprehensively than individual feature-based indicators.

The baseline metrics used for the comparison were the most used and intuitive metrics of Underrepresentation which were:

- Male to female speaker count ratio
- male-to-female voice activity duration ratio
- pitch difference between male and female speakers.
- amplitude and energy differences across channels.

For each baseline, a Kruskal–Wallis H test was performed across the five bias tiers. The results revealed that the count ratio and voice-activity ratio baselines showed moderate discriminatory ability ($\varepsilon^2 \approx 0.39$–$0.42$), indicating that simple gender representation metrics are somewhat sensitive to strong imbalances. However, pitch, amplitude, and energy differences exhibited weak or non-significant separation across tiers ($\varepsilon^2 < 0.18$), demonstrating limited usefulness as bias indicators.

In contrast, EuqAud achieved a substantially larger effect size ($\varepsilon^2 = 0.845$) with strong statistical significance ($p < 4.1 \times 10^{-17}$), outperforming all simple baseline measures by a wide margin. These results confirm that EuqAud captures structured and systematic representational disparities that cannot be detected by trivial heuristics such as raw gender counts or duration ratios. This comparison provides empirical evidence that EuqAud adds explanatory value beyond simple imbalance metrics and strengthens the validity of the metric for real-world gender bias assessment.

Having established that EuqAud outperforms trivial baseline metrics, we next evaluate its performance on diverse real-world audio datasets to confirm its practical utility and robustness.

C. EuqAud ON REAL-WORLD DATASETS

To evaluate how EuqAud works on real world datasets, the best approach, considering the lack of ground truth, was to compare Gender distribution ratio and voice activity (VAD) ratios on real world dataset. Three separate tests were conducted to evaluate EuqAud across diverse real world audio Corpora.

1) TEST 1: TESTING EuqAud PERFORMANCE ON LibriSpeech

The original LibriSpeech dataset provides only a few standard splits (e.g., train, dev, test). To increase the diversity of evaluation, each split was further broken down into random sample subsets of varying sizes. From each subset, the count of audios, Voice activity detection, pitch, amplitude and energy were extracted separately for male and female speakers. These features across all subsets were compiled into structured dataset.

This dataset was passed through EuqAud equation and the Scaling process and then each row was assigned to a tier label based on the EuqAud score the tiers ranged from neutral, moderate and strong. The counts per each tier level was as Neutral: 30, Strong Male-Dominant: 23, Moderate

Male-Dominant: 13, Moderate Female-Dominant: 7, Strong Female-Dominant:3.

Kruskal–Wallis tests confirmed significant differences in both gender ratio dominance ($H = 44.85$, $p < 4.28 \times 10^{-9}$) and voice activity dominance ($H = 42.91$, $p < 1.08 \times 10^{-8}$). This test revealed notable variation across splits in both gender and voice activity ratios, with male-dominant subsets consistently yielding higher EuqAud scores. Overall, the dataset exhibited greater variability compared to subsequent tests, demonstrating EuqAud's sensitivity to representational differences across naturally occurring splits.

2) TEST 2: TESTING EuqAud PERFORMANCE ON COMMON VOICE

The second test used Common Voice datasets, covering multiple language splits and same as Test one the count of audios, Voice activity detection, pitch, amplitude and energy were extracted separately for male and female speakers. Which were then compiled into one structure dataset.

The process of assigning bias tier labels was done to the structure dataset. Counts per bias tier were more evenly distributed, with 56 Neutral, 28 Moderate Female-Dominant, 28 Moderate Male-Dominant, 24 Strong Male-Dominant, and 24 Strong Female-Dominant subsets.

The Kruskal–Wallis tests showed much higher H statistics for both gender ratio ($H = 90.4437$, $p = 1.06 \times 10^{-18}$) and voice activity dominance ($H = 136.2273$, $p = 1.8121 \times 10^{-28}$), reflecting the greater variability across languages.

Compared to LibriSpeech, the Common Voice dataset was more balanced overall, yet EuqAud remained responsive to subtle representational differences, indicating that the metric can detect bias patterns across multilingual datasets.

3) TEST 3: TESTING EuqAud PERFORMANCE ON MULTILINGUAL LibriSpeech

The third test used Multilingual LibriSpeech Datasets which same as Common voice datasets cover multiple languages. The same steps were taken to create the structured dataset and to assign tier labels. The dataset exhibited lower variability, with counts per bias tier of 21 Neutral, 9 Strong Male-Dominant, 7 Moderate Male-Dominant, 4 Moderate Female-Dominant, and 1 Strong Female-Dominant.

Despite this relative homogeneity, Kruskal–Wallis tests still revealed significant differences in both gender ratio ($H = 15.7850$, $p = 0.003322$) and voice activity dominance ($H = 21.1115$, $p = 3.0095 \times 10^{-4}$), demonstrating that EuqAud is sensitive even to subtle bias patterns in more balanced datasets.

Together, these three tests illustrate EuqAud's robustness and consistent responsiveness to representational disparities across datasets of varying sizes, languages, and gender distributions. The tier-based evaluation confirms that the metric can effectively differentiate between datasets with distinct bias profiles.

D. NUMERICAL STABILITY ANALYSIS OF EuqAud

As a manner to evaluate the numerical stability and robustness of EuqAud, we conducted a bootstrap uncertainty analysis across three independently constructed datasets. Each dataset was resampled 1,000 times with replacement, and the mean scaled EuqAud score was recomputed for every bootstrap sample. The process allows to quantify the sensitivity of the metric to dataset perturbations and helps verify whether the directionality and magnitude of detected bias remain stable under resampling.

TABLE 2. Bootstrap stability analysis of EuqAud across three datasets. Mean scaled scores, standard deviation, and 95% confidence intervals demonstrate low variability and stable bias direction under resampling.

Dataset	Mean Scaled EuqAud	95% CI	Std Dev	Interpretation
Dataset 1	0.86	[0.14, 1.83]	0.51	Neutral
Dataset 2	2.21	[0.87, 3.56]	0.7	Mild to Moderate Male Bias
Dataset 3	1.58	[0.57, 2.64]	0.53	Mild Male Bias

As displayed on the table the test done across the three independent datasets, EuqAud was consistent and displayed low variance behavior. Dataset 1 produced a mean scaled EuqAud score of 0.86 with a 95% confidence interval (CI) of [0.14, 1.83], reflecting an overall *balanced* distribution with only minor fluctuations. Dataset 2 yielded a mean of 2.21 (95% CI: [0.87, 3.56]), corresponding to *mild-to-moderate male bias* that remained stable across all bootstrap iterations. Dataset 3 produced a mean score of 1.58 (95% CI: [0.57, 2.64]), indicating *mild male bias* with consistently positive bounds.

Importantly, the confidence intervals for all three datasets remained within the Balanced-to-Moderate range of the EuqAud tier structure. None crossed zero into female bias, and none approached the ± 6 threshold defining Strong Bias. This demonstrates that bias directionality is preserved under perturbation. Standard deviations ranged from 0.51 to 0.70, reflecting low variability and confirming that EuqAud exhibits stable, well-behaved performance even when the underlying dataset is resampled.

These findings collectively validate EuqAud as a reliable and numerically stable metric for dataset-level gender bias estimation.

E. VALIDATING EuqAud'S TIER STRUCTURE

A manually augmented dataset was created by combining feature values extracted from the AMI corpus and TEDLIUM datasets to simulate varying levels of gender bias in a controlled setting. The EuqAud model was then used to compute scores and assign tier labels.

The distribution of labels was approximately balanced across five classes: 20 samples in Strong Male and Strong Female Bias tiers, 44 in each Moderate Bias class, and 40 labeled as Neutral.

A Kruskal–Wallis H test on the EuqAud values across these tiers produced a statistically significant result ($H = 140.12$, $p < 2.7 \times 10^{-29}$), with a large effect size ($\varepsilon^2 = 0.8351$). Dunn's post-hoc test confirmed significant differences between most pairs, validating the tier system's discriminative power.

This test demonstrates that EuqAud scores respond strongly to controlled, manually augmented variations, showing a significant increase in performance when datasets are systematically biased. This provides the confirmation that EuqAud works similarly for both Synthetic and Real-world datasets. To further confirm the validity of the tiering system and the discriminative power of EuqAud in a controlled setting, we conducted tests on a manually augmented dataset with known bias levels.

F. USING EuqAud TO INTERPRET MODEL BEHAVIOR

F1-score differences between male and female speakers were used as a primary measure to assess model level gender performance disparities. The decision to use F1-Score was unlike accuracy, F1-score balances precision and recall, making it robust in the presence of label imbalance. By considering the absolute F1-score difference, the analysis isolates the magnitude of performance disparity without regard to directionality, offering an interpretable proxy for fairness.

To clarify, whereas F1-score is influenced by class distribution, EuqAud measures dataset-level acoustic underrepresentation independent of model outputs. F1-score differences are used here purely as a validation tool to assess whether EuqAud tiers correspond to observable disparities in model performance, not as a direct measure of dataset bias.

To provide empirical evidence for this relationship, a simple neural classifier was trained using MFCC features from the Common Voice Assamese dataset. The architecture included a 40-dimensional input layer, one hidden layer with 64 ReLU units, and an output layer trained with the Adam optimizer. Evaluation metrics included accuracy, F1-score, and confusion matrices. The results confirmed that Across multiple Common Voice datasets, F1-score differences were found to correlate strongly with EuqAud tiers. Datasets labeled as strongly biased according to EuqAud exhibited F1-score differences approaching 1.0, whereas neutral-tier datasets yielded near-zero differences as displayed on the Table 3. As an example, Amharic and Assamese datasets, which had EuqAud above 6, showed stark performance gaps, whereas datasets like Korean and Armenian showed minimal disparities.

These findings demonstrate that the EuqAud and its tier system not only reflect dataset-level imbalance but can also serve as reliable indicators of expected fairness in downstream models.

Table 3 summarizes the relationship between EuqAud scores, tier assignments, and gender-wise F1-scores. While some of the F1-scores in Table 4 take extreme values such as 0 or 1, this behavior is intentional and illustrative rather than indicative of implementation errors. These edge-case scores arise because certain real-world datasets contain audio samples from only a single gender. Including these datasets demonstrates that EuqAud can correctly detect cases of perfect bias, where the representation of one gender is entirely absent. This confirms that EuqAud can capture both moderate and extreme levels of dataset-level gender bias, providing a meaningful and interpretable validation of the metric across the full spectrum of real-world conditions.

TABLE 3. Relationship between EuqAud scores, bias tiers, and absolute gender-wise F1-score differences across selected common voice datasets. Results show that higher EuqAud magnitudes correspond to larger model-level performance disparities.

Common voice dataset	F1_D	EuqAud	Tier
Afrikaans 1	-1	-10	Strong female bias
Amharic	1	10	Strong male bias
Albanian	0.173	7.539	Strong male bias
Assamese	1	6.197	Strong male bias
Asturian	1	6.736	Strong male bias
Hindi	0.042	2.143	Moderate male bias
Korean	-0.01	-0.71	Neutral
Afrikaans 2	-0.01	-0.49	Neutral
Armenian 1	-0.01	-0.89	Neutral
Armenian 2	0.001	0.107	Neutral

Overall, this validation confirms that EuqAud tiering aligns with observed model performance disparities and provides an interpretable, robust measure of gender representation bias in audio datasets.

G. COMPARING EuqAud WITH SYSTEM-LEVEL WER

To validate the effectiveness and reliability of the proposed EuqAud, WER was used as it is a widely accepted metric for evaluating speech recognition performance and has commonly been used to measure disparities in recognition accuracy between male and female speakers.

However, WER reflects the combined influence of both the dataset and the underlying speech recognition model, including factors such as model architecture, pretraining, and optimization strategies. In this study, WER is therefore used only as a validation benchmark for EuqAud. As there are currently no well-established dataset-based validation measures specifically designed for bias detection in audio datasets. Consequently, WER provides a practical point of reference for evaluating the effectiveness of the proposed metric.

In the validation process which yielded the results displayed in Table 4 WER was calculated using the Whisper-tiny speech recognition model developed by OpenAI [44]. Each dataset split was transcribed using this model, and WER values were computed separately for male and female speakers.

TABLE 4. Side-by-side comparison between the WER values for male and female speakers Vs the dataset bias direction inferred from EuqAud. Abbreviations: LBS – LibriSpeech, MLT – Multilingual LibriSpeech, CMV – Common Voice.

Datasets/ Dataset Split	WER male	WER female	System bias defined by WER	Bias score	Dataset bias defined by WER
LBS: Dev	20.05	26.30	Male biased	0.651	Male Biased
LBS: Test-other	16.06	19.06	Male biased	1.263	Male Biased
LBS: Test-clean	22.45	24.75	Male biased	0.811	Male Biased
LBS Train	31.51	32.08	Male biased	0.713	Male Biased
MLT : Portuguese	1.148	1.292	Male biased	-4.790	Female Biased
MLT : Polish	1.222	1.201	Female biased	6.279	Male Biased
CMV : Hakha Chin Train	1.111	1.211	Male biased	1.259	Male Biased
CMV : Hakha Chin other	1.170	1.149	Female biased	-5.241	Female Biased
CMV : Hakha Chin validated	1.323	1.159	Female biased	-0.235	Female Biased
CMV : Chuvash test	1.191	1.226	Female biased	-3.511	Female Biased
CMV : Chuvash validated	1.283	1.296	Male biased	2.512	Male Biased
CMV : Sorani validated split	1.182	1.428	Male biased	8.258	Male Biased
CMV : Western Frisian other	1.499	1.371	Female biased	-4.888	Female Biased
CMV : Western Frisian validated	1.257	1.427	Female biased	-6.172	Female Biased
CMV : Indonesian validated	`1.789	1.479	Female biased	-7.788	Female Biased
CMV : Latgalian validated	1.243	1.452	Female biased	-6.424	Female Biased

EuqAud was used to calculate the biasness of the respective dataset split which WER was calculated for. Table 4 consists of result of the validation done across the two methods.

As displayed in Table 4 In most cases, the bias direction identified by the WER values and EuqAud are aligned, supporting the validity of EuqAud as a metric to identify gender bias in audio datasets. However, a few discrepancies were observed. for example, in the Polish subset of Multilingual LibriSpeech, WER indicates a female bias, whereas EuqAud reveals a strong male bias. Similarly, languages with low overall WER values (e.g., Portuguese, WER < 1.3) still exhibit notable bias scores according to EuqAud.

These observations in Table 4 indicate that there are some discrepancies between EuqAud and WER. These are expected, as WER reflects not only dataset composition but also model-specific factors such as architecture, pretraining, optimization strategies, and error-correction mechanisms as explained by Xu et al. [45] and Jain and Bhowmick [46]. However, this does not completely write out the effect of dataset on the WER. This is a very present fact as explained by Schettino et al., 2025 [47]. But the discrepancy still arises as EuqAud only uses acoustic features such as Pitch and Voice activity extracted from dataset.

Despite these expected differences, for approximately 80% of the dataset splits, the bias directions identified by EuqAud and WER are aligned.

This demonstrates that EuqAud successfully captures gender imbalance independently of model effects using acoustic features from the dataset, while also validating that its assessments correspond closely to system-level performance disparities in most cases.

H. ALIGNMENT BETWEEN EuqAud AND HUMAN PERCEPTION

To assess how well EuqAud aligns with perceived gender representation in speech, a user study was conducted. Participants were presented with 9 short audio clips sampled from a diverse set of Common Voice datasets (e.g., Korean, Afrikaans, Assamese, Albanian, Amharic, Hindi, Armenian, Asturian) representing different tiers of EuqAud scores. Each participant was asked to rate the perceived gender representation of each clip using the following scale: Strong Male Representation, Moderate Male Representation, Neutral, Moderate Female Representation, and Strong Female Representation.

This survey aimed to validate whether the EuqAud metric's tiered outputs align with human perceptions of gender dominance in audio samples, irrespective of language. Participants were instructed to focus on vocal expression attributes (e.g., pitch, tone, vocal presence) rather than content meaning.

Survey responses were aggregated and compared to the tier classifications derived from the EuqAud. Results showed strong agreement between perceptual labels and EuqAud tiers, particularly for strongly biased clips. For example, clips with high positive EuqAud values (e.g., > 6) were predominantly perceived as male-dominant, while highly negative EuqAud values (e.g., < -6) aligned with strong female representation perceptions.

Table 5 summarizes the alignment between the EuqAud determined tier and the dominant perception reported by human participants for each clip. The Agreement column indicates the percentage of participants whose perception matched the tier predicted by EuqAud. High agreement scores across most clips, particularly for extreme bias values, indicate that EuqAud effectively captures bias in a way that is traceable and observable to human listeners. Notably, clips with scores close to zero (e.g., Korean or Armenian) resulted in more mixed human responses, reflecting the subtler nature of perceived neutrality and confirming the sensitivity

TABLE 5. The table displays the similarities between the EuqAud bias tiers and listeners perceived bias for each audio clip, reporting the percentage of participant responses that matched the actual bias level of the provided audio.

Dataset	EuqAud	Tier	Dominant Perception	Agreement
Afrikaans	-10	Strong Female Bias	Strong Female Representation	91.90%
Amharic	10	Strong Male Bias	Strong Male Representation	89.20%
Assamese	6.078	Strong Male Bias	Strong Male Representation	86.50%
Asturian	6.551	Strong Male Bias	Strong Male Representation	94.60%
Hindi	-2.758	Moderate Female Bias	Strong Female Representation	54.10%
Armenian	-1.56	Neutral	Neutral	83.80%
Indonesian	-2.09	Moderate Female Bias	Neutral	70.30%
Korean	-1.912	Neutral	Moderate Female Representation	56.80%
Albanian	4.944	Moderate Male Bias	Moderate Male Representation	86.50%

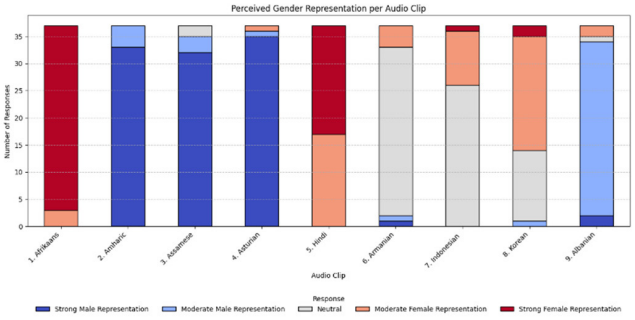

FIGURE 3. Illustrates the distribution of participant responses for each audio clip, showing how listeners perceived the speaker's gender. Responses are categorized into five perception levels such as Strong male, Moderate male, Neutral, Moderate female, and strong female biased audio clips, highlighting the variation in gender assumptions across the provided audio samples.

of EuqAud in detecting gradations of bias. While Figure 3 complements this by visually depicting the distribution of participant responses per audio clip, reinforcing the consistency between EuqAud tiers and perceptual judgments across the five gender representation categories.

These findings suggest that EuqAud not only captures statistical and performance-based disparities, but also aligns with real-world perceptual judgments, further supporting its utility as a model-independent and traceable fairness metric.

V. DISCUSSION

Whereas the proposed EuqAud 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. In such scenarios, EuqAud could potentially be combined with automatic gender prediction models to infer speaker labels prior to bias computation. However, this integration introduces additional risks because gender classifiers themselves may contain demographic bias or domain-specific errors. Incorrect gender predictions could propagate into EuqAud calculations, producing misleading bias estimates. Therefore, when automatic gender inference is used, the reliability of the gender classification system must be carefully evaluated, and uncertainty estimates should be considered when interpreting EuqAud scores.

Audio datasets used in this study include multiple speakers; each audio clip includes a single speaker. 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. Furthermore, the datasets used in this study primarily consist of clean, read, single-speaker, near-field speech. EuqAud has not yet been evaluated on far-field audio, noisy environments, overlapping speakers, or child/elderly speech, and its generalizability to these conditions remains an open direction for future work.

One critical insight from our extended validation is the challenge of detecting female biased datasets in real world corpora. Most publicly available datasets exhibited male dominance in either sample count or voice activity. This imbalance skews model performance and limits the representativeness of fairness evaluations. Future data collection efforts must prioritize demographic balance to enable equitable AI development and bias mitigation.

Additionally, the EuqAud regression model was trained using synthetic bias scores generated by systematically reducing one gender's acoustic features in controlled increments. While this approach allowed the model to learn a predictable mapping between feature differences and bias scores, it may not fully reflect all possible patterns of natural bias in real-world datasets. Although the regression model was trained using controlled synthetic distortions, EuqAud was subsequently validated on multiple real-world audio corpora, demonstrating that the metric generalizes beyond the artificial training conditions and is sensitive to naturally occurring gender imbalances. Nevertheless, synthetic distortions cannot fully capture the complexity of natural demographic imbalances, and EuqAud's sensitivity to more diverse real-world distributions will require future validation using datasets with verified demographic ground truth. Also, as EuqAud uses a fixed trained equation without retraining on target datasets, characterizing coefficient stability through uncertainty estimation (e.g., bootstrap resampling, confidence intervals, or cross-domain evaluation) becomes

especially important for confirming robustness across new environments.

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.

VI. CONCLUSION

This study introduces EuqAud, a metric designed to quantify gender bias in audio datasets. By leveraging raw audio features and avoiding dependency on downstream models or algorithms, EuqAud offers a dataset focused perspective on bias detection

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 EuqAud and model-based bias indicators. Importantly, the proposed metric isolates dataset based bias, offering a critical advantage over traditional system level metrics like WER.

EuqAud 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-Aty, M. Hafez, S. Alkhazaleh, M. A. Deif, and R. ELGohary, "Parkinson's disease detection based on vocal biomarkers and machine learning approach," in *Proc. Int. Telecommun. Conf. (ITC-Egypt)*, Cairo, Egypt, Jul. 2024, pp. 475–480.

[2] H. Ko and S. Kwon, "Optimization of voice biomarkers to predict Alzheimer's disease," *Alzheimer's Dementia*, vol. 19, no. S15, p. 79907, Dec. 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," *Interdiscipl. Nursing Res.*, vol. 3, no. 3, pp. 190–198, 2024.

[4] I. Michels, V. Urovi, H. Strik, H. A. C. V. Helvoort, S. O. Simons, and I. L. Michels, "Vocal biomarkers in COPD: Capturing disease severity using voice," *Eur. Respiratory J.*, vol. 60, p. 1591, Mar. 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 *Proc. Inter-Noise Noise-Con Congr. Conf.*, Chiba, Japan, 2023, pp. 859–870.

[6] R. Jahangir, Y. W. 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 Syst. Appl.*, vol. 171, Jun. 2021, Art. no. 114591.

[7] D. Dablain, B. Krawczyk, and N. Chawla, "Towards a holistic view of bias in machine learning: Bridging algorithmic fairness and imbalanced learning," *Discover Data*, vol. 2, no. 1, p. 18, Apr. 2024.

[8] E. Ferrara, "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," *Science*, vol. 6, no. 1, p. 3, Dec. 2023.

[9] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. 1st Conf. Fairness, Accountability Transparency*, 2018, pp. 77–91.

[10] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Proc. Equity Access Algorithms, Mech., Optim.*, New York, NY, USA, Oct. 2021, pp. 1–9.

[11] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 33–44.

[12] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, and K. Crawford, "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, Dec. 2021.

[13] B. Hutchinson, A. Smart, A. Hanna, R. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell, "Towards accountability for machine learning datasets: Practices from software engineering and infrastructure," in *Proc. 21st Conf. Fairness, Accountability Transparency*, 2021, pp. 560–575.

[14] 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," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 14, pp. 7684–7689, Apr. 2020.

[15] G. Attanasio, B. Savoldi, D. Fucci, and D. Hovy, "Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2024, pp. 21318–21340.

[16] 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," *Comput. Speech Lang.*, vol. 60, Mar. 2020, Art. no. 101032.

[17] W. Hutiri, T. Patel, A. Y. Ding, and O. Scharenborg, "As biased as you measure: Methodological pitfalls of bias evaluations in speaker verification research," in *Proc. Interspeech*, Kos, Greece, Sep. 2024, pp. 4268–4272.

[18] J. Meyer, L. Rauchenstein, J. D. Eisenberg, and N. Howell, "Artie Bias corpus: An open dataset for detecting demographic bias in speech applications," in *Proc. 12th Lang. Resour. Eval. Conf.*, Marseille, France, 2020, pp. 6462–6468.

[19] S. Feng, O. Kudina, B. Mark Halpern, and O. Scharenborg, "Quantifying Bias in automatic speech recognition," 2021, *arXiv:2103.15122*.

[20] M. Zanon Boito, L. Besacier, N. Tomashenko, and Y. Estève, "A study of gender impact in self-supervised models for speech-to-text systems," in *Proc. Interspeech*, Sep. 2022, pp. 1278–1282.

[21] F. Burkhardt, B. T. Atmaja, A. Derington, F. Eyben, and B. Schuller, "Check your audio data: Nkululeko for bias detection," in *Proc. 27th Conf. Oriental COCOSDA Int. Committee Co-ordination Standardisation Speech Databases Assessment Techn. (O-COCOSDA)*, Hsinchu, Taiwan, Oct. 2024, pp. 1–6.

[22] M. Shabbir, A. Hussain, and M. M. Khan, "Age and gender estimation through speech: A comparison of various techniques," in *Proc. 18th Int. Conf. Emerg. Technol. (ICET)*, Peshawar, Pakistan, Nov. 2023, pp. 228–233.

[23] A. Ghosal, C. Pathak, P. Singh, and S. Dutta, "Voice-based gender identification using co-occurrence-based features," in *Proc. Comput. Intell. Pattern Recognit.*, 2019, pp. 947–956.

[24] E. Priya, S. J. Priyadharshini, P. S. Reshma, and S. Sashaank, "Temporal and spectral features based gender recognition from audio signals," in *Proc. Int. Conf. Commun., Comput. Internet Things*, Chennai, India, 2022, pp. 23–29.

[25] Y. M. Ali, E. Noorsal, N. F. Mokhtar, S. Z. M. Saad, M. H. Abdullah, and L. C. Chin, "Speech-based gender recognition using linear prediction and mel-frequency cepstral coefficients," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 28, no. 2, p. 753, Nov. 2022.

[26] S. Fahmeeda, M. A. Ayan, M. Shamsuddin, and A. Amreen, "Voice based gender recognition using deep learning," *Int. J. Innov. Res. Growth*, vol. 3, pp. 649–654, Feb. 2022.

[27] S. Srivastava, H. Sharma, and D. Garg, "Comparative study of machine learning algorithms for voice based gender identification," in *Proc. Int. Conf. Edge Comput. Appl. (ICECAA)*, Tamilnadu, India, Oct. 2022, pp. 1136–1141.

[28] Z. Zhang, R. Li, and K. Chen, "Speaker gender recognition based on semi-supervised learning," in *Proc. 3rd Int. Conf. Comput., Commun., Perception Quantum Technol. (CCPQT)*, Zhuhai, China, Oct. 2024, pp. 232–235.

[29] D. Ika A, E. Hartati, and R. Karw, "An analysis of intonation pattern in the pre service English teacher's talks," *FRASA, English Educ. Literature J.*, vol. 3, no. 2, pp. 64–71, Sep. 2022.

[30] S. Ekeruke, "Stress patterns and intonations among the annang language speaking students of faculty of arts, Akwa Ibom state university," *J. Cmmunication Culture*, vol. 12, no. 3, pp. 163–172, 2024.
[31] B. Ludusan, M. Heldner, and M. Wlodarczak, "Exploring the role of formant frequencies in the classification of phonation type," in *Proc. Int. Congress Phonetic Sci.*, 2023, pp. 1726–1730.
[32] A. Bailey and M. D. Plumbley, "Gender bias in depression detection using audio features," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Dublin, Ireland, Aug. 2021, pp. 596–600.
[33] 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 *Proc. 12th Lang. Resour. Eval. Conf.*, Marseille, France, 2020, pp. 4218–4222
[34] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, South Brisbane, QLD, Australia, Apr. 2015, pp. 5206–5210.
[35] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 2757–2761.
[36] A. Rousseau, P. Deleglise, and Y. Esteve, "TED-LIUM: An automatic speech recognition dedicated corpus," in *Proc. 8th Int. Conf. Lang. Resour. Eval.*, Istanbul, Turkey, 2012, pp. 125–129.
[37] 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 *Proc. 6th Int. Conf. Lang. Resour. Eval.*, Marrakech, Morocco, 2008, pp. 137–140.
[38] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 5582–5586.
[39] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using Kaldi," in *Proc. Interspeech*, 2017, pp. 498–502.
[40] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," 2014, *arXiv:1412.5567*.
[41] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 7, pp. 1315–1329, Jul. 2016.
[42] B. Cao, K. Teplansky, N. Sebkhi, A. Bhavsar, O. Inan, R. Samlan, T. Mau, and J. Wang, "Data augmentation for end-to-end silent speech recognition for laryngectomees," in *Proc. Interspeech*, Incheon, South Korea, Sep. 2022, pp. 3653–3657.
[43] C. Kim, M. Shin, A. Garg, and D. Gowda, "Improved vocal tract length perturbation for a state-of-the-art end-to-end speech recognition system," in *Proc. Interspeech*, Sep. 2019, pp. 739–743.
[44] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Mach. Learn.*, 2023, pp. 28492–28518.
[45] J. Xu, Z. Yang, A. Zeyer, E. Beck, R. Schlueter, and H. Ney, "Dynamic acoustic model architecture optimization in training for ASR," in *Proc. Interspeech*, 2025, pp. 3603–3607.
[46] P. Jain and A. Bhowmick, "Comparative performance analysis of end-to-end ASR models on indo-aryan and dravidian languages within India's linguistic landscape," *EURASIP J. Audio, Speech, Music Process.*, vol. 2025, no. 1, p. 10, Feb. 2025.
[47] L. Schettino, V. Vitale, and A. Vietti, "Toward optimised datasets to fine tune ASR systems leveraging less but more informative speech," Cagliari, Italy, 2025, pp. 1048–1054.

SRIWANTHI JAYAWARDENA received the B.Sc. degree (Hons.) in information technology, specializing in data science from Sri Lanka Institute of Information Technology (SLIIT), Malabe, Sri Lanka. Her research interests include machine learning, responsible artificial intelligence, bias detection in AI systems, large language models, and deepfake detection. She is particularly interested in developing trustworthy and transparent AI systems through improved data representations and evaluation methodologies.

PRASANNA S. HADDELA received the B.Sc. and M.Sc. degrees in computer science from the University of Colombo, Sri Lanka, and the Ph.D. degree from Sheffield Hallam University, U.K.

He was a Visiting Lecturer at the Open University of Sri Lanka, for more than 15 years. He is currently a Lecturer and a Research Supervisor at Global Curtin, Curtin University, Australia. He is a Senior Lecturer with the Faculty of Computing, Sri Lanka Institute of Information Technology, where he is the Dean/Academic of Matara Centre and the Head of Graduate Studies-FoC. He is also a Research Fellow at INTI International University, Malaysia. He has authored more than 60 international publications and has actively contributed to the organization of international research conferences. His research interests include explainable AI, text analytics, and neurocomputing.

Dr. Haddela is a fellow of the Higher Education Academy, U.K. He also serves as a reviewer for several international journals and conferences.

THISARA SHYAMALEE received the M.Sc. (by research) degree from the University of Moratuwa, Sri Lanka. She is currently a Lecturer with the Faculty of Computing, Sri Lanka Institute of Information Technology. Her research interests include machine learning, deep learning, explainable AI, and human–computer interaction.

AMANDI EKANAYAKE received the B.Sc. degree (Hons.) in information technology, specializing in data science from Sri Lanka Institute of Information Technology, Malabe, Sri Lanka. Her research interests include machine learning, responsible AI, and explainable AI.

THARUSHI MUDALIGE received the B.Sc. degree (Hons.) in information technology, specializing in data science from Sri Lanka Institute of Information Technology, Malabe, Sri Lanka. Her research interests include machine learning, responsible AI, and scene embeddings.

IMESHA DHANAWARDHANA received the B.Sc. degree (Hons.) in information technology, specializing in data science from Sri Lanka Institute of Information Technology, Malabe, Sri Lanka. Her research interests include machine learning, responsible AI, and word embeddings.