AUDIO BIAS SCORE: BIAS DETECTION METRIC FOR GENDER BIAS DETECTION IN AUDIO DATASETS.

K. M. S. P. Jayawardena

(IT21380914)

BSc (Hons) Degree in Information Technology
Specializing in Data Science

Department of Computer Science

Sri Lanka Institute of Information Technology
Sri Lanka

April 2025

AUDIO BIAS SCORE: BIAS DETECTION METRIC FOR GENDER BIAS DETECTION IN AUDIO DATASETS.

K. M. S. P. Jayawardena

(IT21380914)

Dissertation submitted in partial fulfilment of the requirements for the Bachelor of Science (Hons) Degree in Information Technology Specializing in Data Science

Department of Computer Science

Sri Lanka Institute of Information Technology
Sri Lanka

April 2025

DECLARATION.

“I declare that this is my own work and this dissertation1 does not incorporate without acknowledgement any material previously submitted for a degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to Sri Lanka Institute of Information Technology, the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:  Date: 2025.04.11

The supervisor/s should certify the dissertation with the following declaration.

The above candidate has carried out research for the bachelor’s degree

Dissertation under my supervision.

Signature of the supervisor: Date:

ABSTRACT.

With the advancement of audio-based Artificial Intelligence (AI) systems in critical domains such as healthcare and law enforcement, the need for fairness and responsibility is becoming increasingly important, particularly regarding gender-related bias. One prevalent challenge in ensuring fairness is the ability to quantify and interpret gender bias within a system. Prior studies have addressed this issue by introducing metrics that primarily focus on measuring performance bias, rather than addressing biases present in the underlying dataset. To address this gap the study introduces a new metric built from raw audio features such as pitch, energy, and amplitude along with voice activity. The metric is built in a way where the speakers' language, race or age would not affect its performance. Polynomial regression with L2 regularization was used to build the bias quantification equation. The equations' performance was checked across multiple datasets where it's r2 values was always within the range of 95% to 99% .The metric will provide a score ranging between 10 to -10 where when a dataset is unbiased the score is provided as 0 and when the datasets is 100% biased towards males it provides a score of 10 and when the dataset 100% biased toward females it provides a score of -10. As this metric is built in a language independent way and as it focusses on raw audio features, this metric can be applied to any audio dataset which contains the metadata of the speaker gender.

Keywords: gender bias, audio data, artificial intelligence, responsible AI

ACKNOWLEGEMENT.

I would like to express my sincere gratitude to my supervisor Dr. Prasanna S. Haddela and co-supervisor Thisara Shayamalee for their continuous support, encouragement and valued guidance throughout the course of this research. Without their guidance the successful completion of the research would have been impossible. Especially, during the challenging phases of the project where navigating through complex problems were made possible by their expertise.

I would also like to extend my gratitude to my fellow group members, E. M. A. M. Ekanayake, H. M. I. K. Danawardana, and T. N. Mudalige for assisting me, providing thoughtful discussions and ideas during each phase of the research.

I am grateful to all the academic and non-academic staff of S. L. I. I. T. who assisted me in various ways be it academic, administrative or technical support. A special thanks is extended to the Librarian at S. L. I. I. T. who always assisted me with accessing necessary resources.

TABLE OF CONTENTS

LIST OF FIGURES.

LIST OF TABLES.

LIST OF ABBREVIATIONS.

Abbreviation	Meaning
AI	Artificial Intelligence
PWC	PricewaterhouseCoopers
FPR	False Positive Rate
FNR	False Negative Rate
EER	Equal Error Rate
WER	Word Error Rate
CER	Character Error Rate
MSE	Mean Squared Error
NMSE	Normalized Mean Squared Error
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error

1. INTRODUCTION.

1.1 Background.

The advancement of Artificial Intelligence (AI) has been significant. With this advancement, the use of artificial intelligence in critical fields for decision making has increased as well. These systems use different modals of data when training AI based systems to make decisions. One such modal is audio data, apart from system such as voice assistants, Audio based AI models are used in Critical functionalities such as predicting Parkinson's [1], Alzheimer's [2] and neurodegenerative [3] diseases [4], search and rescue operations [5] and forensic voice analysis [6] used in law enforcement.

The increase of usage of these systems in critical systems means the need for more responsibility and fairness. It is important that these systems do not reflect any bias in them. To understand the word bias there are multiple definitions provided by leading AI developers such as Google, PWC and Azure. The definition provided by PWC states that an AI system is said to be biased, when AI systems make decisions that are systematically unfair to certain groups of people [7]. It is found that this systematic unfairness is caused by training data being unrepresentative or reflecting existing prejudices [8], from the design and structure of the AI algorithms [7] and by interactions of the AI systems with the real world [8].

The research focuses on quantifying the bias that is introduced by training data being unrepresentative and the training data reflecting existing prejudices against a gender. The study is built on such biases found in Audio datasets, which can be quantified by the metric that is built.

1.2 Literature Survey.

Bias measures are used to quantify bias as a way of recognizing if a system or a dataset is biased towards a certain group. Throughout the years there have been multiple approaches to measure bias that is found in audio based AI systems. Most of such approaches use statistical base metrics such as False positive rates (FPR), false negative rates (FNR), Equal Error rate (EER) and Minimum detection cost(minCDet) [9]. These performance based metrics help quantify decision making errors by comparing system's behavior for different demographic groups. FPR and FNR measure rates at which model incorrectly classify positive and negative instances. While EER point out across which groups the FPR and FNR values are equal. The minimum detection cost combines the error rates and assign weights to the cost of false positives and negatives.

With the statistical methods as the base performance metrics, few other customized methods such as G2mindiff, G2avgratio, G2avg log ratio are used to detect bias in audio based systems, where G2mindiff calculate the base performance metric (EER/FPR/FNR…) of the base group and the group which has the best base performance metric (EER/FPR/FNR…) . G2avg ratio takes the base performance metric(EER/FPR/FNR…) of the group you want to check the system's biasness and the average base performance metric across the different demographic groups provides a ratio, which can be used to find out relative performance disparities of different demographic groups. G2avg log ratio takes the log value of G2avg ratio [10] providing a scaled perspective on the performance difference between different demographic groups.

$$G2mindiff = \ b_g - b_m$$

- b_g: $base\ metric(EER, FPR, FNR)\ for\ selected\ demographic\ (g)$
- b_m: $base\ metric(EER, FPR, FNR)\ for\ best\ performing\ demographic\ (m)$

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

- b_g: $base\ metric(EER, FPR, FNR)\ for\ selected\ demographic\ (g)$

- $\bar{b}$: $Average\ base\ metric\ across\ all\ demographic\ groups$

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

Apart from the methods mentioned, measures such as Character error rate(CER) [11] and Word Error rate(WER) [12] [13] are commonly employed in audio based AI systems. Especially when detecting bias in Speech recognition and language processing models. Where CER quantifies difference between predicted and actual characters in transcription, WER measures the difference between predicted and actual words in transcription.

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

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

While existing measures quantify bias or fairness of a system primarily by evaluating model outcomes such as error rates or classification accuracy, it is important to acknowledge that the performance bias is a combination of bias introduced by the dataset which is used to build the model and underlying algorithm. These metrics do not account for the bias introduced within the dataset itself, where factors such as imbalance demographic representation or poor quality audios can introduce significant and often overlooked bias.

As current bias detection metrics in audio-based AI systems largely focus on performance indicators, the need for a metric that exclusively targets inherent bias found in audio data itself remains. The metrics introduced in the study aim to address this gap by focusing on bias found in the audio data independent of the model performance.

1.3. Research Gap.

Existing bias detection metrics for audio-based AI systems primarily focus on model performance, such as error rates and classification accuracy, without considering the bias introduced during dataset creation. These performance-based metrics fail to address the inherent biases present within the dataset itself. Specifically, class imbalances, the quality of audio data, and other dataset-related factors that contribute to bias are often overlooked. This research aims to bridge this gap by developing a metric that focuses solely on the dataset level, isolating gender bias introduced by the training data, independent of the model's performance

1.4. Research Problem.

The core problem addressed by this research is the absence of effective tools for detecting gender bias introduced at the dataset level in audio-based AI systems. Current bias detection techniques predominantly assess the fairness and bias of AI systems based on performance outcomes, such as false positive and false negative rates, or the accuracy of model predictions. While these methods are valuable for evaluating model fairness, they do not capture the biases that stem from the training datasets themselves. Gender bias in audio datasets, whether due to unbalanced demographic representation, inadequate data quality, or other factors, can lead to the development of biased AI systems that unfairly impact certain groups.

This research problem is particularly pertinent as gender bias in audio data can exacerbate issues related to gender inequality, affecting AI applications in critical domains such as healthcare, law enforcement, and emergency services. Therefore, there is an urgent need for a dedicated metric that can isolate and measure the gender bias embedded in the dataset, independent of any model performance factors. By addressing this gap, the research will offer a methodology to quantify gender bias at the dataset level, providing a foundation for the development of fairer and more equitable audio-based AI systems.

1.5. Research Objectives.

1.5.1. Main objective.

The main objective of this study is to design and implement a metric that specifically quantifies gender bias present in audio datasets, focusing on raw audio features. The metric will be formulated to isolate dataset-level biases, providing a clear understanding of how gender disparities manifest within the dataset before any model intervention.

1.5.2. Sub-objectives.

- Understanding how different audio features can cause bias.
  The purpose of the objective is to explore how various raw audio features, such as pitch, amplitude, speech rate, and energy, can be influenced by gender, potentially leading to bias in the dataset. It will assess how these features might cause disparities in speech recognition or other audio-based AI tasks and identify which features are most critical for quantifying gender bias.
- Understanding which audio features are affected by the language and race of the speaker.
  In this objective it will be investigated how different languages and the race of the speaker impact audio features. As the metric is built in a language-independent way it is important to understand which audio features are effected by speaker's race or the language and if it could adjusted in a way that it would not hinder the performance of the metric when applied to datasets containing different languages.

2. METHODOLOGY.

2.1. Selection Of Features For Equation Building.

The selection of features which were focused on when building the equation were done through studying previous research based on audio-based gender classification systems by analysing which features were considered when building such systems. Through the research studies the primary features considered were found to be pitch [14], Amplitude [15], Energy [16], Formants, Intonations and Mel-Frequency Cepstral Coefficients (MFCCs) [17] [18] [19] [20].

Intonations [21] [22] and formants [23] are dependent on the language and vary among cultures and ethnic groups of the speaker as the metric is built independent of the language and race with the sole focus on identifying Gender bias in datasets these features were not considered when building the metrics. Also, it is stated by Bailey Et al. (2021) [24] that models based on raw audio are more robust to gender biased than ones based on hand-crafted features, such as mel-spectrograms [24]. Therefore, raw audio-based features were only focused on when building the metric.

Additionally, Number of Audios and voice activity per gender is used in the metrics. Where number of audios are considered as a way for identifying class imbalances and Voice activity to quantify even if the classes among genders are properly balanced with the difference of voice activity it can affect the performance of a model trained on such a dataset.

Gender based variations naturally exist in the factors pitch, Amplitude and Energy levels, where pitch is inherently higher for female speakers. Energy levels and amplitude also display consistent high or low values for a specific gender depending on the speech characteristic. For the quantified bias to not be affected by these inherent behaviours the standard deviations of the pitch, amplitude and energy levels were used.

In conclusion, the count of audio files per gender, Voice activity per gender, Standard deviation of Amplitudes, Energy and pitch per gender was used to build the equation.

2.1.1. Selected features explained.

I. Pitch

Pitch is one of the most fundamental characteristics of human speech and is directly associated with the fundamental frequency produced by the vibration of the vocal cords. In audio signals, pitch is perceived as the frequency of a sound wave and is typically measured in Hertz(Hz). It serves as a powerful and widely used indicator for gender classification systems, this is mainly due to the physiological differences between male and female speakers. Male speaker generally possess thicker and longer vocal folds which vibrate at a lower frequency range which is approximated to fall between the range of 85Hz to 180Hz . In contrast, female speakers tend to have a shorter and thinner vocal fold that vibrate at a higher frequency.

However, pitch is not solely related to the speakers' gender, it is also modulated by the emotional state, linguistic context and individual speech habits. These factors introduce natural variation within gender groups which, if not accounted for, could misinterpret pitch based measurements .

To avoid conflating natural biological difference with potential systematic bias, this study utilizes the standard deviation of pitch with each gender group rather than sum or the mean of the pitch values. This approach enable the identification of inconsistencies in pitch distribution across genders that are reflective of datasets construction bias rater than innate vocal characteristics.

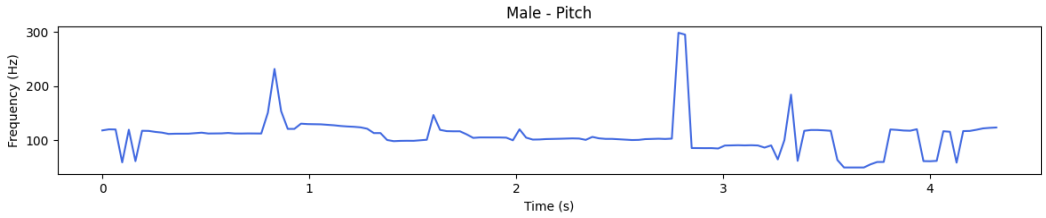

Figure 1 Single male speaker pitch distribution over time.

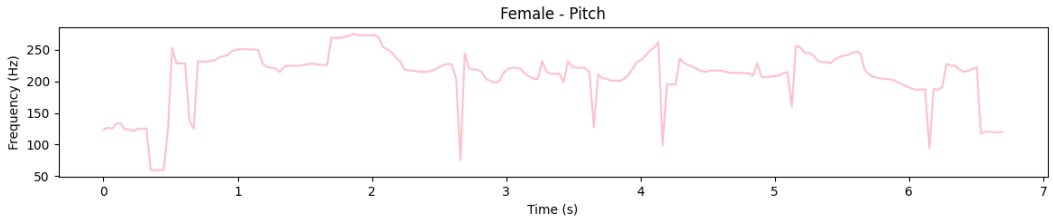

Figure 2 Single female speaker pitch distribution over time.

II. Amplitude.

Amplitude reflects the strength or magnitude of and audio signal, corresponding to the level of air pressure fluctuations produced during speech. In waveform representation, amplitude is depicted by the height of the wave and is closely related to the perceived loudness or intensity of speech. While, higher amplitude values generally denote louder speech, the amplitude of an audio signal can be influenced by several factors, including speaking style, vocal effort and environmental conditions.

Physiologically, amplitude differences between genders may occur due to varying subglottal pressures and vocal fold dynamics. However, amplitude is also susceptible to systematic factors such as inconsistencies in microphone placement, room acoustics and speaker posture. These factors can disproportionately affect one gender over another, especially in datasets compiled from diverse sources or uncontrolled recording environments.

In order to address these concerns the study uses the standard deviation of amplitude per gender, rather than relying on absolute amplitude values. This statistical measure provides insights into the internal variability of amplitude within each gender, which offers a assessment of how loudly or softly speakers are represented in the dataset in specific gender.

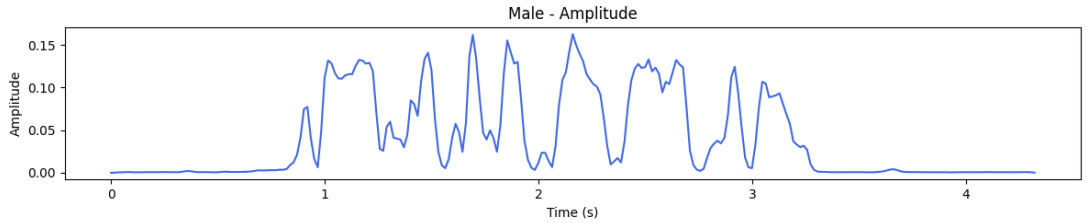

Figure 3 Single male speaker amplitude distribution over time.

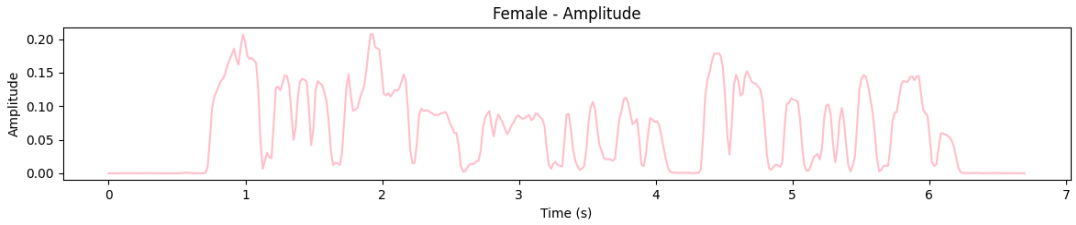

Figure 4 Single female speaker amplitude distribution over time.

III. Energy.

Energy in an audio signal is a measure of total power of the sound wave over time. It is typically calculated by squaring the amplitude values of a signal within a given ti,e window and summing them to obtain cumulative acoustic power. This feature captures the force or effort behind speech production and plays a significant role in applications such as speaker activity detection, emotion analysis and speech clarity assessments.

The energy level may vary due to lung capacity, vocal effort and articulation habits. However, similar to amplitude energy is influenced by physiological as well as technical and environmental variables. Recording inconsistencies microphone sensitivity etc. If one gender is disproportionately affected by such external variables the dataset may exhibit bias.

Similar to pitch and amplitude in order to avoid the influence of these confounding factors, standard deviation of energy is calculated separately for each gender in this study. Using standard deviation allow to evaluate how uniformly energy is distributed within a gender.

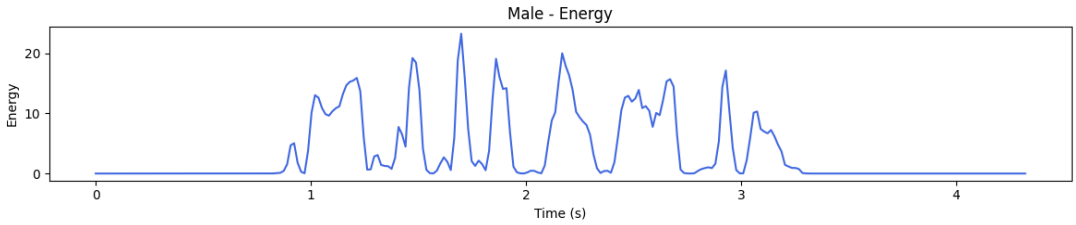

Figure 5 Single male speaker energy distribution over time.

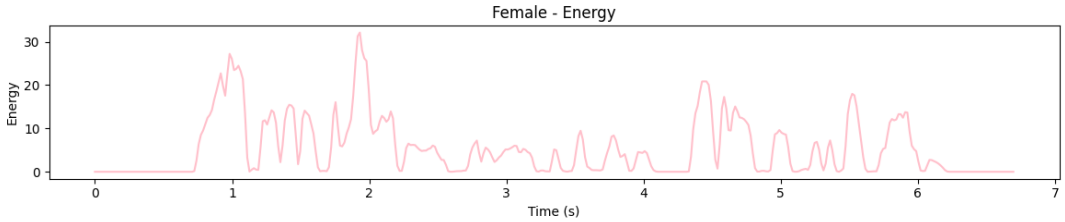

Figure 6 Single female speaker energy distribution over time.

IV. Voice activity per gender.

Voice activity is defined as the temporal duration during which a speaker is actively producing speech excluding periods of silence, background noise or non-verbal sounds. This metric is crucial for assessing not just the presence of the audio in samples in a dataset but the actual contribution of each gender in terms of usable content-rich speech.

In many audios datasets equality in the number of audio files per gender does not necessarily equate to equality in representation. While male and female speaker maybe present in equal numbers, if one gender consistently speaks longer durations per audio clip the datasets will inadvertently favor that group in terms of speech content. This discrepancy can influence model performance by providing more training data and phonetic diversity for one gender leading to biased learning outcomes.

To account for this voice activity per gender is computed to quantify the actual spoken duration across all samples within each gender category. By incorporating this metric, the study ensures that bias detection is not limited to numerical class balance but extends to qualitative aspects of speech as well.

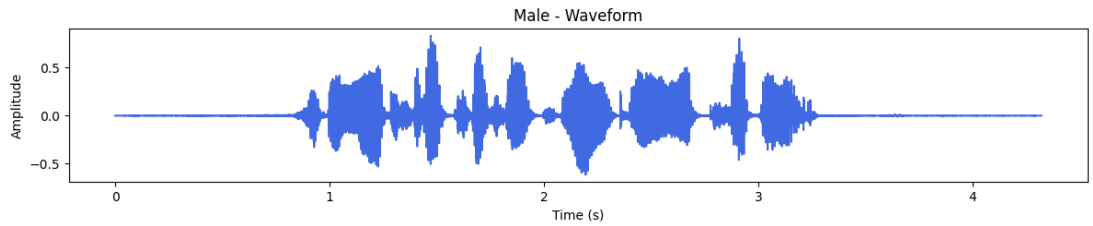

Figure 7 Single male speaker waveform

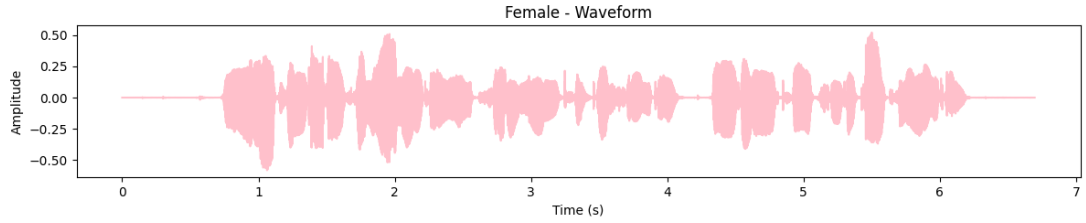

Figure 8 Single female speaker waveform

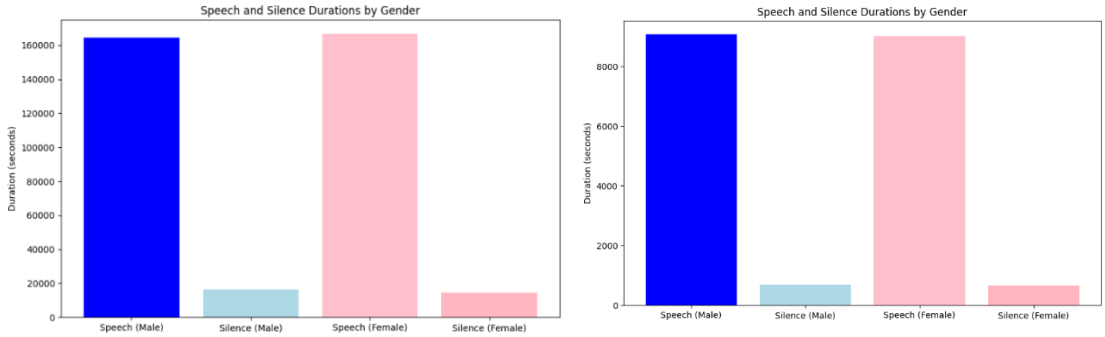

Figure 9 Voice Activity of Genders in LibriSpeech 360 (Left) and LibriSpeech 100 (Right)

V. Number of audio samples per gender.

The count of audio samples per gender is a fundamental measure of identifying class imbalance in speech datasets. And equal number of recordings between male and female speaker can directly affect the performance of the machine learning models training on such data. Class imbalance often leads to biased model predictions favoring the overrepresented class while under performing on the minority class. This class imbalance not only affects classification accuracy but also fairness in real-world deployment.

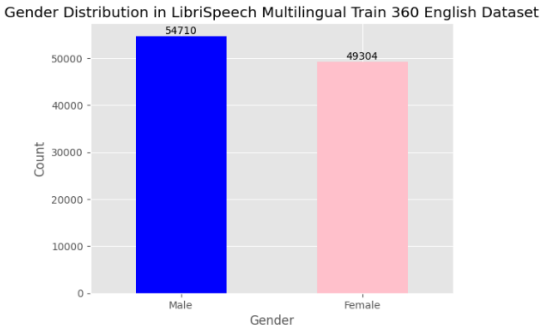

Figure 10 Gender Distribution LibriSpeech Multilingual Train 360

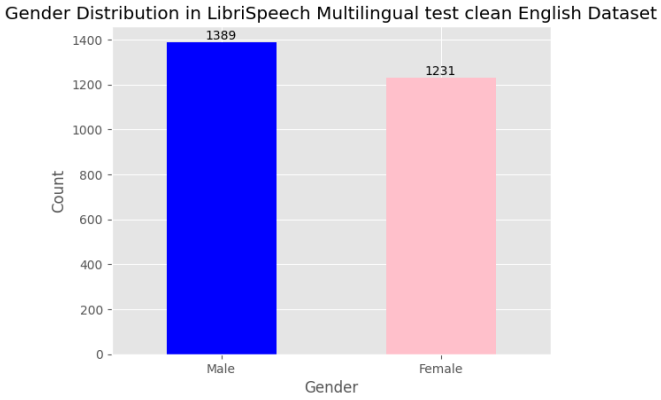

Figure 11 Gender Distribution LibriSpeech Multilingual Test dataset

2.2. Building the dataset.

2.2.1. Introducing the datasets.

I. Common Voice Dataset.

The Mozilla Common Voice dataset [25] is a large-scale, open-source collection of voice recordings contributed by volunteers around the world. It was created to help train machine learning models for automatic speech recognition (ASR) and other speech-related tasks. As of its latest versions, Common Voice includes millions of samples across more than 90 languages and dialects, making it one of the most linguistically diverse open-source speech corpora available.

Each audio sample is paired with its corresponding transcription, and the dataset is structured with detailed metadata including speaker gender, age, accent, and

recording device information. This rich set of metadata makes Common Voice especially useful in bias analysis and fairness evaluation across demographic variables. Due to the crowd-sourced nature of data collection, it exhibits natural variations in recording quality, pronunciation, pitch, and speech speed, which are valuable in simulating real-world conditions. Its diverse linguistic representation and demographic labeling make it an ideal candidate for building a base dataset for gender bias evaluation in audio data.

II. LibriSpeech Dataset.

The LibriSpeech dataset [26] is one of the most well-known corpora in the field of speech processing. It is derived from audiobooks that are part of the public domain LibriVox project. The recordings are primarily in English and are accompanied by time-aligned text transcriptions. The dataset is divided into various subsets based on quality and speaker characteristics, such as "train-clean-100", "train-clean-360", and "train-other-500".

LibriSpeech offers high-quality, 16kHz recordings and includes over 1,000 hours of speech. Its structure provides speaker labels and gender information, making it suitable for demographic analysis. Because the speakers are reading text in a controlled manner, the recordings are generally of high clarity and consistent linguistic content, which is advantageous for extracting clean acoustic features such as pitch, amplitude, and energy. It has been widely used as a benchmark dataset in both traditional and deep learning-based speech recognition research.

III. LibriSpeech Multilingual dataset.

The LibriSpeech Multilingual (LibriVox Multilingual Speech) [27] dataset is an extension of the original LibriSpeech, aimed at promoting research in multilingual ASR systems. It contains recordings from the LibriVox project in multiple languages including French, German, Spanish, and Italian, among others. Just like its predecessor, it provides audiobooks read by volunteers, but this time in several different languages, each with aligned text transcriptions.

This dataset is essential for studying language-independent speech characteristics. It allows researchers to investigate how acoustic features vary across languages and whether bias detection models trained on language-agnostic features remain effective. Because it includes gender and speaker ID metadata, it supports comparative gender bias studies across languages. It's particularly useful for developing and validating language-neutral metrics, as intended in this research.

IV. TED-LIUM Dataset.

The TED-LIUM dataset [28] is constructed from TED talks, which are presentations delivered by speakers from around the world on a broad range of topics. The dataset includes transcriptions aligned with audio at the word level. It features diverse speakers with varying accents, intonations, speech styles, and language proficiencies, offering a realistic and challenging set of conditions for speech recognition and bias analysis.

What sets TED-LIUM apart is the variability in speech caused by spontaneous delivery, public speaking nuances, and emotional expression. Unlike read speech in LibriSpeech, TED-LIUM features natural, conversational language. This introduces fluctuations in pitch, energy, and speaking rate—core components used in your bias metric. The dataset includes speaker metadata such as gender and speaker identity, allowing researchers to evaluate how models generalize across demographic and expressive variances.

V. AMI meeting corpus.

The AMI (Augmented Multi-party Interaction) Meeting Corpus [29]is a dataset that captures real-world meeting conversations among multiple participants. It consists of approximately 100 hours of meeting recordings, including both audio and video, transcriptions, dialogue acts, and metadata on speaker roles and demographics.

The audio in AMI is characterized by overlapping speech, background noise, and informal conversational dynamics, reflecting highly realistic communication

scenarios. The corpus was recorded using multiple microphone types (individual headset mics and room mics), introducing variations in audio quality. This diversity in acoustic environments and speaker interactions makes it a valuable resource for testing the robustness of audio features such as energy distribution and voice activity under more complex, real-world conditions.

The availability of speaker role, gender, and conversation style information also makes AMI ideal for testing how gender bias manifests not just in solo speech but in interactive settings. This dataset is particularly useful for examining how bias scores may be influenced by group dynamics or speech interruptions, adding depth to the validation process of the proposed metric.

Combination of Each of these datasets' different splits as well as subsets for different languages were used to build one training and testing dataset.

	A	B	C	D	E	F	G	H	I	J	K
1	Dataset split	Count_male	Count_female	voice_activity_male	voice_activity_female	energy_male	energy_frmale	amplitude_male	amplitude_female	pitch_male	pitch_female
2	Librispeech Englsih Dataset										
3	Dev-clean(337MB)	1374	1329	8946.75	9020.14	1282.84	1197.81	450.94	708.26	42.25	45.56
4	Dev-other(314MB)	1450	1414	8468.01	8319.99	1289.06	1203.7	387.43	409.11	38.62	44.26
5	test-clean(346MB)	1389	1231	9073.79	9013.66	1188.67	1119.96	376.01	376.68	44.34	55.15
6	test-other 328	1561	1378	8651.95	8708.53	1298.58	1067.23	511.5	335.28	46.32	46.7
7	train clean 100 6.3GB	14342	14197	164431	166619	1189.73	1116.82	616.51	432.79	46.08	47.62
8	train clean 360 23GB	54710	49304	619310	574195	1146.69	1188.46	528.36	655.45	46.84	48.15
9	Multiligual Librispeech Dataset										
10	Italian_train	32609	27014	434400	362925	1341.03	1022.05	984.58	549.15	38.38	47.36
11	Italian_test	694	568	9207.42	7893.26	1040.45	1012.51	643.98	357.28	35.54	34.33
12	italian_dev	544	704	7596.22	9516.49	1365.35	954.68	772.14	403.68	33.93	38.68
13	portuguese_train	12865	24668	166937.9	327847.42	988.97	790.62	334.57	285.12	40.22	52.17
14	portuguese_test	447	424	6201.42	5975.38	939.32	833.45	295.7	238.28	34.46	40.03
15	portuguese_dev	508	318	7405.32	4677.85	948.29	786.19	275.72	306.25	58.76	57.08
16	polish_train	18691	6352	262854.4	91507.09	844.49	851.29	242.28	320.39	45.25	63.27
17	polish_test	255	265	3479.44	3556.55	744.2	1110.75	216.71	386.9	42.02	47.08
18	polish_dev	229	283	3175.68	3750.11	1135.5	767.06	290.79	190.69	46.75	39.22
19	Common voice Dataset										
20	arabic_train	3612	4624	13299.15	16075.8	2259.04	2636.67	744.12	1166.16	22.54	19.64
21	br_test	776	209	1906.84	517.66	1905.71	1785.44	1146.51	1041.81	20.47	18.21
22	br_validation	862	157	1981.46	407.88	1722.29	1951.06	937.62	686.23	21.97	27.73
23	cnh_train	192	163	567.32	435.25	1356.7	1257.75	418.67	665.36	25.88	11.35
24	cnh_other	1176	815	2646.93	2583.07	624.48	1451.89	848.41	969.38	17.75	15.37
25	cnh_validated	608	467	2045.78	1421.24	1300.81	1385.5	849.2	911.84	19.29	15.25
26	cv_train	1046	105	4177.5	427.36	1129.06	2418.14	1340.83	478.76	25.14	22.44
27	cv_test	324	105	1396.79	449.87	2340.06	3595.41	1192.15	907.62	23.82	18.65
28	cv_validated	9366	4816	39611.78	19852.82	1731.49	2754.49	1087.17	733.02	20.7	21.76
29	cy_train	3510	2659	14884.86	11492.82	1978.5	2238.04	736.05	963.01	22.24	24.68
30	cy_validation	1538	1246	6689.72	5683.38	2176.02	1853.45	827.53	944.25	23.6	23.98
31	cy_test	693	595	3043.32	2587.22	1868.41	1678.06	942.1	970.84	24.35	29.38
32	cy_other	6215	3789	26737.4	16813.16	1834.56	1653.12	919.3	920.27	23.36	31.34
33	cy_validated	36205	27229	141901.4	112437.26	1935.57	1993.45	1004.07	1052.17	21.67	23.62

Figure 12 Extracted Features stored in CSV.

2.2.2. Processing and building the Dataset.

Unlike classification tasks that have clearly defined labels (e.g., male vs. female), bias detection lacks an inherent target variable. There is no predefined value that indicates how biased a sample or dataset is unless an external benchmark or metric is imposed. Since the features considered in the proposed metric, such as standard deviations of

acoustic attributes are relatively novel in this context, no existing benchmark bias scores were available for supervised training or evaluation. To address this the creation of a synthetic bias score needed to be done.

In order to create a synthetic bias score a controlled data augmentation needed to be performed on the existing extracted values. The idea of the controlled data augmentation approach was to simulate bias by systematically altering feature values associated with one gender group while keeping the other unchanged, and vice versa. This allowed for the controlled introduction of disparities between genders along the selected features, thereby creating a gradient of bias that could be quantitatively tracked.

The augmentation process involved the following steps:

- The curated dataset was duplicated, resulting in two equal-sized subsets—each containing the same audio files and corresponding feature vectors.
- In the first half of the augmented dataset, feature values for male speakers were kept unchanged, while feature values for female speakers were reduced in controlled increments of 5%. This reduction was applied to all core features. For example, a feature value of 0.20 for a female speaker would be modified to 0.19 (5% reduction), 0.18 (10% reduction), and so forth across different versions of the sample.
- In the second half of the dataset, the process was reversed: female speaker features were kept constant, while male speaker features were incrementally reduced in the same 5% steps.
- Each modification introduced a known level of imbalance between the two gender groups, allowing a bias score to be associated with each sample. This score reflected the degree of artificial bias introduced via feature manipulation. For instance, an unmodified sample was assigned a score of 0.0, while a sample

with a 5% reduction in one gender's features was assigned a score of 0.05, with the scale extending to higher values such as 0.10, 0.15, and so on.

- The final augmented dataset thus contained samples with associated bias scores ranging from 0.0 to 1.0, representing a linear and interpretable scale of bias intensity introduced through controlled feature perturbations.

	A	B	C	D	E	F	G	H	I	J	K
1	count_mal	count_fem	voice_activ	voice_activ	energy_ma	energy_fen	amplitude_	amplitude_	pitch_male	pitch_fema	score
2	693	34.65	50.722	2.5361	1868.41	93.4205	942.1	47.105	24.35	1.2175	4.75
3	6215	310.75	445.6233	22.28117	1834.56	91.728	919.3	45.965	23.36	1.168	4.75
4	36205	1810.25	2365.023	118.2512	1935.57	96.7785	1004.07	50.2035	21.67	1.0835	4.75
5	71487	3574.35	3574.59	178.7295	2100.74	105.037	983.69	49.1845	25.58	1.279	4.75
6	24650	1232.5	1363.977	68.19883	1942.62	97.131	910.38	45.519	26.62	1.331	4.75
7	6237	311.85	358.2067	17.91033	1330.55	66.5275	808	40.4	28.55	1.4275	4.75
8	8448	422.4	623.6887	31.18443	1653.63	82.6815	941.48	47.074	25.2	1.26	4.75
9	4455	222.75	291.18	14.559	1640.67	82.0335	1181.14	59.057	26.2	1.31	4.75
10	12505	625.25	691.9583	34.59792	1827.01	91.3505	994.97	49.7485	21.45	1.0725	4.75
11	4304	215.2	239.1013	11.95507	1450.97	72.5485	834.85	41.7425	29.93	1.4965	4.75
12	13006	650.3	1282.221	64.11103	1176.53	58.8265	1031.47	51.5735	23.47	1.1735	4.75
13	4643	464.3	304.4187	30.44187	1734.59	173.459	789.55	78.955	26.61	2.661	4.5
14	14814	1481.4	909.0247	90.90247	1598.69	159.869	756.15	75.615	28.78	2.878	4.5
15	11029	1102.9	735.8437	73.58437	1797.3	179.73	924.51	92.451	28.4	2.84	4.5
16	11441	1144.1	954.402	95.4402	1708.79	170.879	955.86	95.586	35.1	3.51	4.5
17	15555	1555.5	1052.119	105.2119	1748.45	174.845	947.13	94.713	28.01	2.801	4.5
18	6708	670.8	53.704	5.3704	2031.87	203.187	1471.13	147.113	23.77	2.377	4.5
19	20530	2053	289.991	28.9991	1465.55	146.555	935.74	93.574	28.41	2.841	4.5
20	15500	1550	1060.504	106.0504	1409.54	140.954	892.18	89.218	28.14	2.814	4.5
21	5814	581.4	420.549	42.0549	1386.29	138.629	870.94	87.094	28.73	2.873	4.5
22	12373	1237.3	125.738	12.5738	2029.45	202.945	7544.28	754.428	21.88	2.188	4.5
23	13525	1352.5	760.981	76.0981	1482.32	148.232	1080.08	108.008	26.07	2.607	4.5
24	22418	3362.7	1239.951	185.9927	1834.95	275.2425	1104.5	165.675	45.62	6.843	4.25
25	14418	2162.7	936.6322	140.4948	1504.41	225.6615	1115.3	167.295	35.76	5.364	4.25
26	2781	417.15	145.901	21.88515	1831.28	274.692	1050.33	157.5495	47.27	7.0905	4.25
27	15720	2358	988.0373	148.2056	1957.85	293.6775	1163.17	174.4755	23.38	3.507	4.25

Figure 13 Dataset with controlled Data Augmentation Technique Applied.

2.3. Building And Training The Equation.

The training dataset was created by applying the controlled data augmentation technique to base dataset. As the intended method of developing a predictive equation was linear regression, it was compulsory to assess whether the dataset met the key assumptions required for its application. Therefore, the dataset was evaluated on the assumption's linearity, independence of observations, Homoscedasticity, normality of residuals and absence of multicollinearity.

2.3.1. Explaining the assumptions of Linear regression.

I. Linearity.

Linearity refers to the assumption that the relationship between the dependent variable (Y) and the independent variables (X_1, X_2, ..., X_n) is linear in parameters. That is, changes in the input features lead to proportional changes in the output. The standard linear regression model is represented by the equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

- Y: Dependent variable (bias score)
- $X_1, X_2, \ldots, X_n$: Independent variables (e.g., standard deviations of audio features)
- β_0: Intercept
- $\beta_1, \ldots, \beta_n$: Coefficients of the predictors
- ε: Error term (residual)

To verify this assumption, scatter plots were initially used to visually inspect the relationships between each predictor and the response variable. Additionally, Pearson's correlation coefficient and the correlation matrix were utilized as quantitative measures to assess the strength and direction of linear associations between pairs of variables. These tools collectively ensured that the linearity condition was sufficiently met.

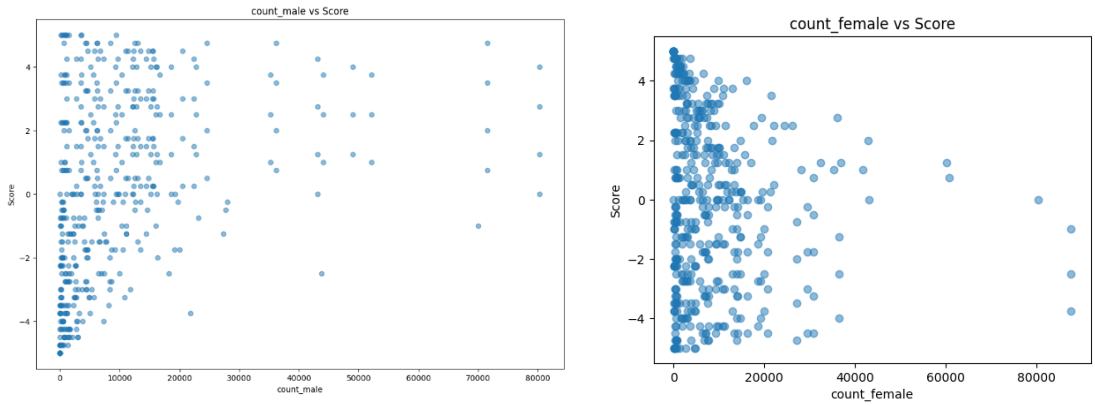

Figure 14 Linearity check : Count Vs Score - Male count(Left), Female Count (Right)

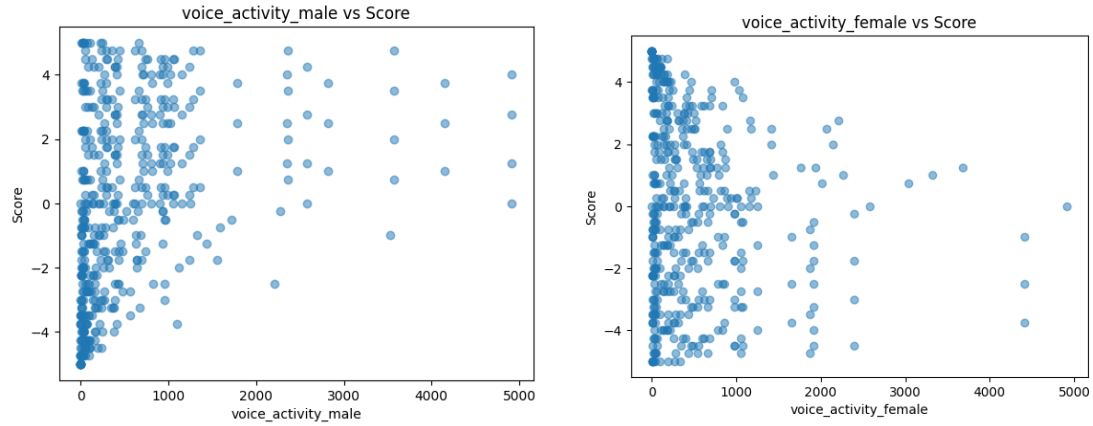

Figure 15 Linearity check : Voice Activity Vs Score - Male Voice Activity(Left), Female Voice Activity (Right)

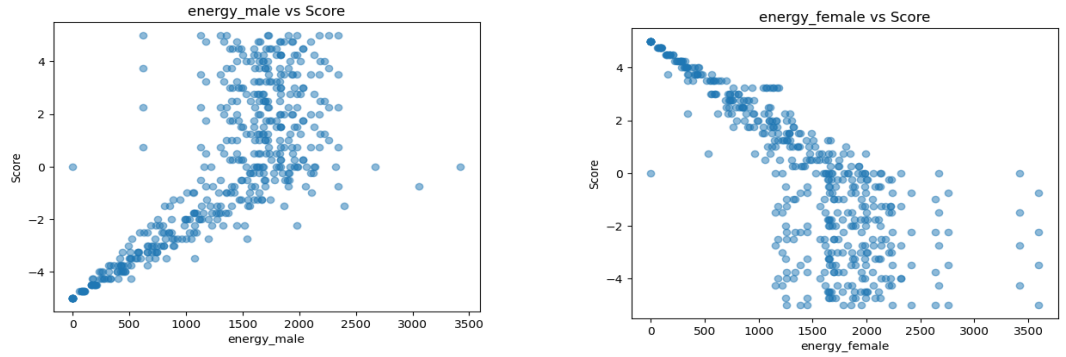

Figure 16 Linearity check : Std. Energy Vs Score - Male Std. Energy(Left), Female Std. Energy (Right)

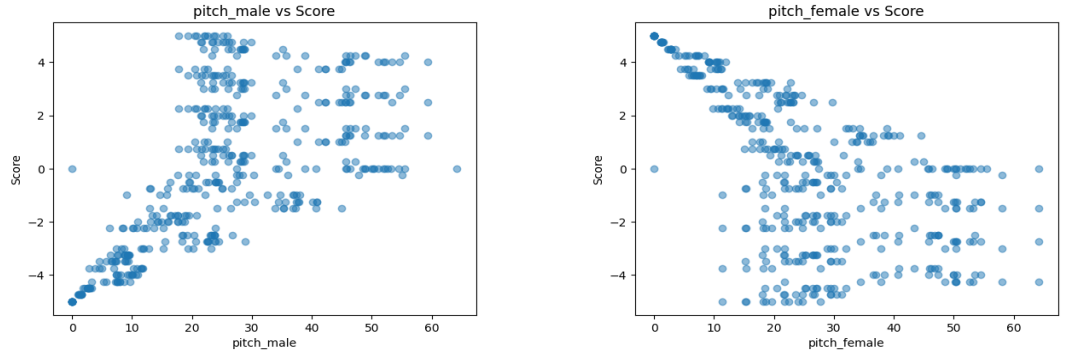

Figure 17 Linearity check : Std. Pitch Vs Score - Male Std. Pitch(Left), Female Std. Pitch (Right)

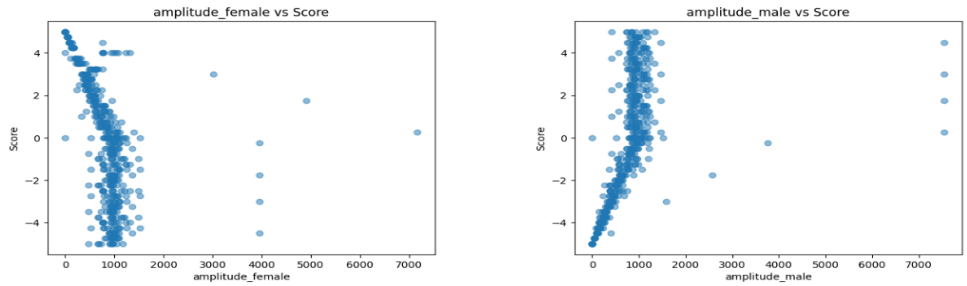

Figure 18 Linearity check : Std. Amplitude Vs Score - Male Std. Amplitude(Left), Female Std. Amplitude (Right)

Pearson Correlation with Score:

	Feature	Pearson Correlation
4	energy_male	0.753126
8	pitch_male	0.628183
6	amplitude_male	0.426615
0	count_male	0.342894
2	voice_activity_male	0.332102
1	count_female	-0.131895
3	voice_activity_female	-0.138098
7	amplitude_female	-0.445231
9	pitch_female	-0.604615
5	energy_female	-0.785568

Figure 19 Pearson's Correlation Score

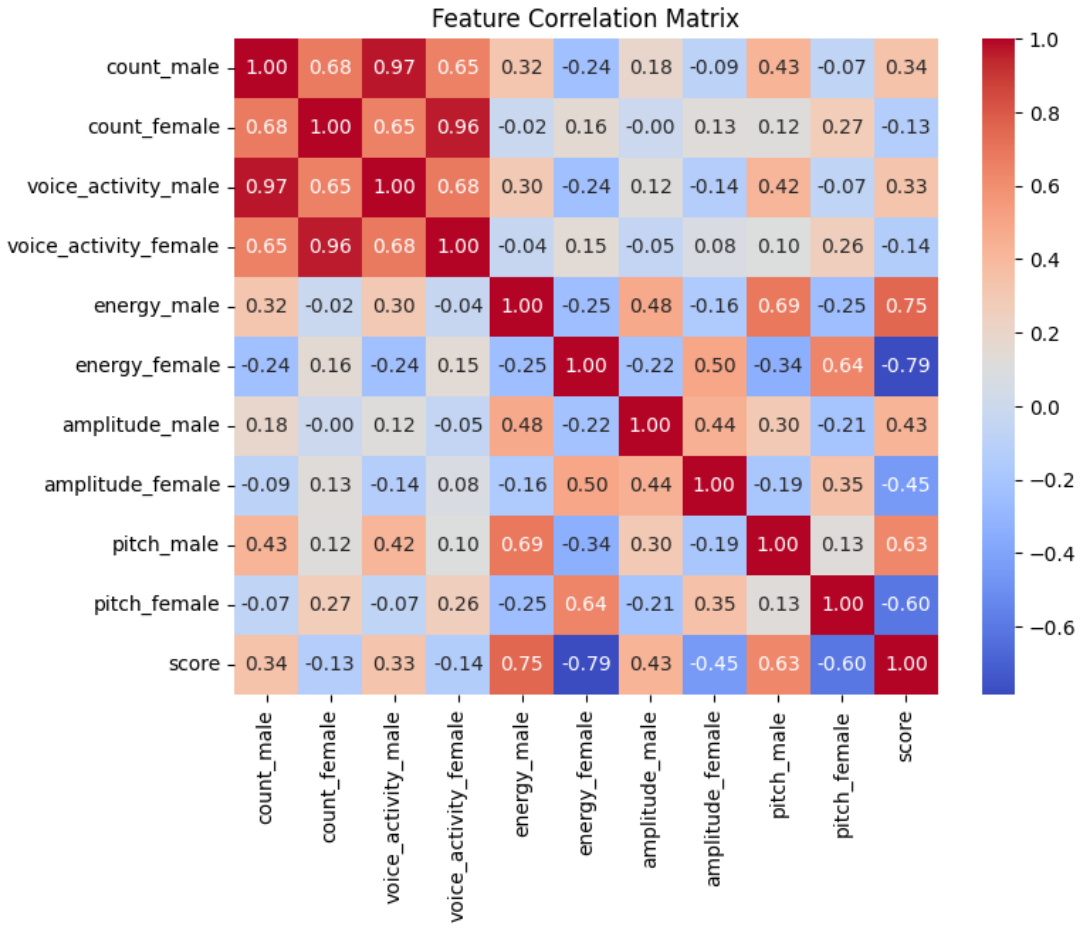

Figure 20 Correlation matrix

II. Independence of Observations.

This assumption states that each observation in the dataset should be independent of the others, meaning that the residuals (errors) of the model should not be correlated across observations. This is particularly critical in time series or sequential data, where autocorrelation may exist. The violation of this assumption can result in underestimated standard errors, leading to overconfident inferences.
To assess this, the Durbin-Watson statistic was computed, defined as:

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Where:

- e_t: residual at time *t*,

Values of DW close to 2 indicate no autocorrelation, whereas values below 1 or above 3 suggest positive or negative autocorrelation respectively. The datasets' Durbin-Watson statistic was equal to the value of 1.0276 displaying a possible positive autocorrelation.

III. Homoscedasticity.

Homoscedasticity refers to the requirement that the residuals exhibit constant variance across all levels of the independent variables. When this condition is violated the standard errors of the regression coefficients become unreliable, which affects hypothesis testing.
This assumption was evaluated by plotting the residuals versus the fitted values. A random scatter of points around the horizontal axis with no discernible pattern is indicative of homoscedasticity. Furthermore, formal statistical tests such as the Breusch-Pagan test and White's test was employed to detect the presence of non-constant variance in the residuals.
In this study, the visual inspection of residual plots did not reveal any funnel-shaped patterns, thereby confirming the homoscedasticity assumption.

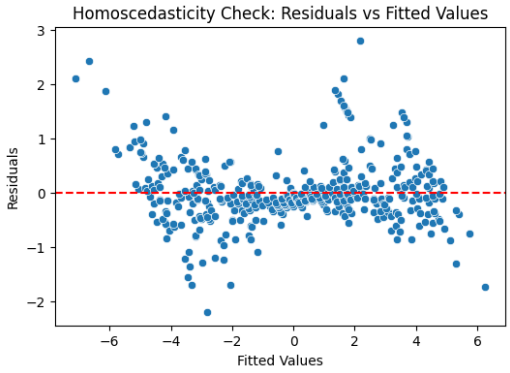

Figure 21 Homoscedasticity check

Apart from the visual inspections the statistical tests Breusch-Pagan test and White's test displayed the following values respectively

- Breusch-Pagan test
    - LM stat: 82.087
    - LM p-value: 1.95×10^{-13}
    - F-stat: 9.741
    - F p-value: 8.54×10^{-15}

  Given that the p-values for both the LM statistic and F-statistic are significantly less than the conventional threshold of 0.05,the null hypothesis was rejected suggesting that heteroscedasticity is present.
- White's test
    - White's test statistic: 210.34
    - Degrees of freedom: 88
    - p-value: 5.12×10^{-12}

  Given the extremely low p-value, the null hypothesis was rejected, concluding that heteroscedasticity is present in the model residuals.

IV. Normality of Residuals.

The normality assumption stipulates that the residuals of the regression model should be normally distributed. This assumption is essential to ensure the validity of confidence intervals and hypothesis tests concerning the regression coefficients. To

evaluate this, a Q-Q (Quantile-Quantile) plot was generated, which plots the quantiles of the residuals against the theoretical quantiles of a standard normal distribution. If the residuals are normally distributed, the points on the Q-Q plot should align closely with the 45-degree diagonal line.

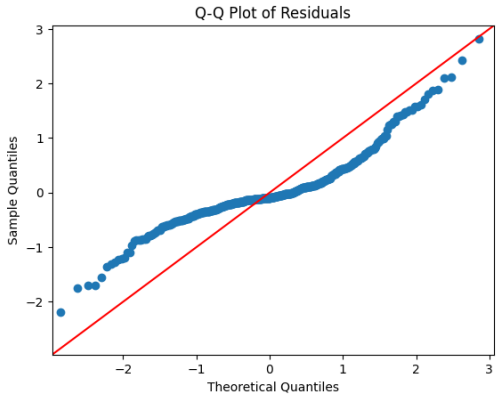

Figure 22 Q-Q Plot of residuals

In addition to the Q-Q plot, the Shapiro-Wilk test and the Kolmogorov-Smirnov test was used for statistical validation of normality and they displayed the results as follows:

- Shapiro-Wilk Test
  - Test Statistic: 0.957
  - P-value: 2.41×10^{-10}

  Given that the p-value is significantly smaller than the conventional threshold of 0.05, we reject the null hypothesis (H_0), indicating that the data is not normally distributed.

- Kolmogorov-Smirnov Test.
  - Test Statistic: 0.293
  - P-value: 1.33×10^{-35}

  Given that the p-value is extremely small (much smaller than the threshold of 0.05), we reject the null hypothesis (H_0), concluding that the data does not follow a normal distribution.

V. Absence of Multicollinearity.

Multicollinearity occurs when two or more independent variables are highly correlated with each other, leading to inflated standard errors and unreliable estimates of regression coefficients. The Variance Inflation Factor (VIF) was employed to diagnose multicollinearity among the predictors, calculated as:

$$VIF_i = \frac{1}{1 - R_I^2}$$

Where:

- R_I^2: coefficient of determination.

Variance Inflation Factor (VIF):

	Feature	VIF
0	const	13.486425
1	count_male	41.965076
2	count_female	35.075759
3	voice_activity_male	42.444733
4	voice_activity_female	35.483454
5	energy_male	4.354491
6	energy_female	4.304774
7	amplitude_male	3.281165
8	amplitude_female	3.228986
9	pitch_male	4.993634
10	pitch_female	4.089872

Figure 23 Variance Inflation Factor

The dataset violated Normality, Multicollinearity, Homoscedasticity, and Independence of the observations. Therefore, simple linear regression could not be used for equation building. To build the equation in a way where these violations will not affect the performance of the metric the choice of methods Symbolic regression and Polynomial regression with L2 (Ridge) regularization were available. Each of these methods were used to build an equation after which the performance was compared, and a final decision was made based on the performance of each equation.

2.3.2. Building the equation: Method - Symbolic regression.

The training dataset was utilized to develop a predictive model using symbolic regression, an evolutionary algorithm-based technique that searches the space of mathematical expressions to find an optimal equation that best fits the data. To optimize the model's performance, a hyperparameter tuning process was carried out

by experimenting with different configurations. The hyperparameters explored included:

- Population size: [500, 1000, 1500]
- Number of generations: [5, 10, 15]
- Crossover probability (p_crossover): [0.6, 0.7]
- Point mutation probability (p_point_mutation): [0.05, 0.1]

Following extensive experimentation, the best-performing configuration was found to be:

- Population size: 500
- Point mutation probability: 0.05
- Crossover probability: 0.7
- Generations: 15

This combination produced the lowest Mean Squared Error (MSE) value of 6.64160, indicating good predictive accuracy. The symbolic regression algorithm subsequently generated a closed-form mathematical expression that models the relationship between the selected input features and the target variable as follows:

$$A = (0.460 - x_6) * ((x_7 - x_6) * (x_9 * x_7))$$

$$B = ((((-0.015 - x_4) * (-0.776 + X_5)) - \frac{X_6}{X_8}) + X_6$$

$$C = (-0.015 - X_4) + \left(\frac{X_6}{(X_1 + X_5)}\right) + 0.959 - \left(\frac{X_5}{(X_{8*}(X_4 + X_5))}\right) - ((x_9 + x_6) + \left(\frac{x_1 + (x_9 - x_3)}{X_2} * X_7\right)) + \left(\frac{X_1 * (X_0 - 0.781)}{X_8}\right) + (\frac{-0.752}{-0.055})$$

$$D = \frac{x_8}{x_4}$$

$$E = x_9 + (((-0.015 - x_4) * (-0.776 + x_5) * x_9) + \left((x_7 - x_6) * (x_9 * x_7)\right))$$

$$F = -\left(\frac{x_6}{\left(\left((x_3 - x_8) * (x_1 + x_5)\right) * x_7 + \frac{x_5}{x_9}\right)}\right) - x_1$$

$$Score = A + B + C + D + E + F$$

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*

Upon extensive evaluation of the bias score generated by the metric, a notable pattern emerged: datasets biased towards male speakers consistently produced positive bias scores, while datasets biased towards female speakers yielded negative scores. This directional sensitivity of the metric was sufficient to indicate the presence and direction of gender bias within a given dataset. However, while the score effectively signaled *whether* a bias existed, it did not inherently quantify the magnitude of that bias in a meaningful or interpretable range.

To address this, the bias score was normalized using min-max scaling, transforming the values to fall within a fixed range of -10 to 10, where +10 would ideally represent complete male bias, -10 complete female bias, and 0 a balanced dataset. This scaling was intended to enhance interpretability and provide a consistent measure of the extent of bias in any given dataset.

However, further testing revealed a limitation in this approach. The scaled scores were accurate and meaningful only up to a bias level of approximately 40% in either direction. Beyond this threshold, the score began to gradually decrease, rather than continuing to increase in alignment with the introduced bias. These findings are visually represented in the graphs provided below, which illustrate the degradation in scoring consistency beyond the 40% bias point.

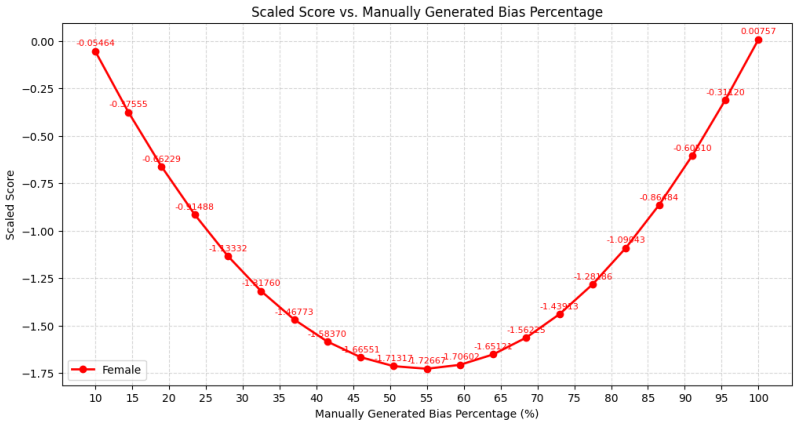

Figure 24 Method- Symbolic Regression : Female Scaled Score Vs Generated bias percentage

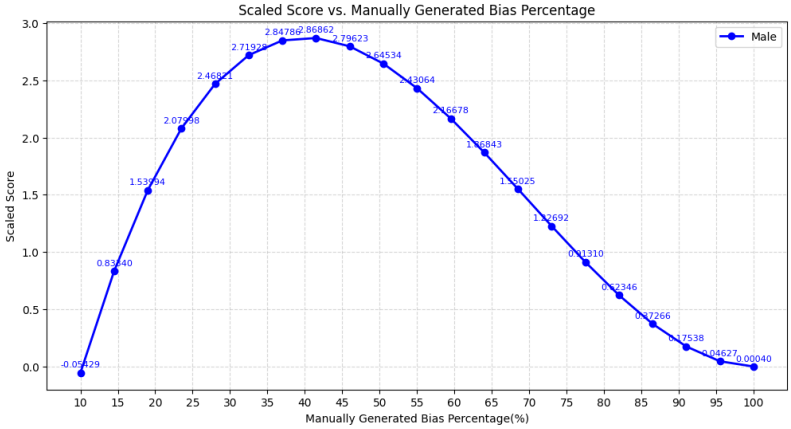

Figure 25 Method- Symbolic Regression : Male Scaled Score Vs Generated bias percentage

As it does not perform the way it is expected to and as the MSE value is significantly higher A new methodology was needed to be studied.

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

The augmented dataset was further utilized to train a predictive model using the Elastic Net Regression technique. Elastic Net combines both L1 (Lasso) and L2 (Ridge) regularization, making it well-suited for scenarios where multicollinearity exists and where feature selection is beneficial. To identify the optimal configuration, the model was trained iteratively using a grid search approach across a range of hyperparameters. Specifically, the regularization strength (alpha) was varied over the set [0.001,0.01,0.1,1.0,10.0][0.001, 0.01, 0.1, 1.0, 10.0][0.001,0.01,0.1,1.0,10.0], while the mixing parameter (l1_ratio) was varied over [0.0,0.1,0.5,0.9,1.0][0.0, 0.1, 0.5, 0.9, 1.0][0.0,0.1,0.5,0.9,1.0], where a value of 0.0 corresponds to pure Ridge regression and 1.0 corresponds to pure Lasso regression.

The best performing configuration was found to be:

- Alpha: 0.01
- L1 Ratio: 0.0 (indicating a purely Ridge-based solution)

This configuration yielded a Mean Squared Error (MSE) of 0.0016 and an R-squared value (R^2) of 0.9998, suggesting a nearly perfect fit to the data with very low prediction error. The final regression equation derived from this model is as follows:

$$Bias\ Score = -0.0125 + (0.0001C_{male}) + (-0.0001C_{female}) + (0.0005V_{male}) + (-0.0005V_{female}) + (0.0044E_{male}) + (-0.0034E_{female}) + (-0.0004A_{male}) + (-0.0004A_{female}) + (-0.0334P_{male}) + (-0.0325P_{female}) + (0.0002P_{male}^2) + (-0.0002P_{male}P_{female}) + (0.0002P_{female}^2$$

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}: $Standard\ deviation\ of\ Pitch\ female$

An extensive testing process was done to analyze how the equation performed on different datasets like with symbolic regression it was found out that when a dataset is biased towards males the score is a positive value and when the dataset is biased towards female the score is a negative value.

Figure 26 Method- Polynomial Regression : Female Scaled Score Vs Generated bias percentage

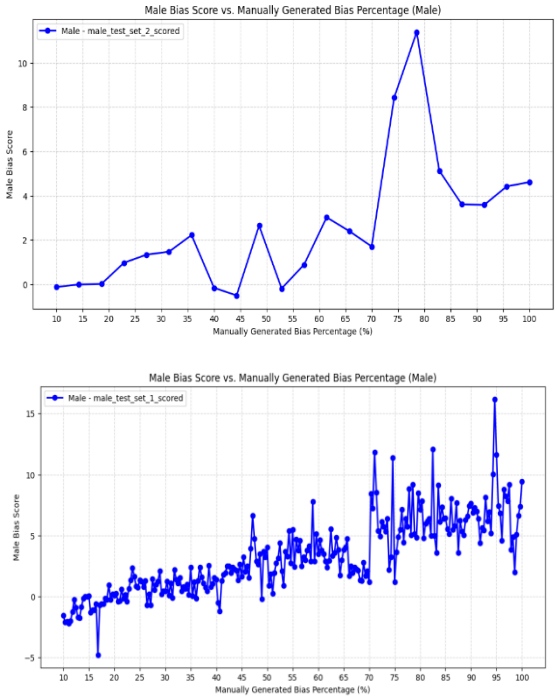

Figure 27 Method- Polynomial Regression : Male Scaled Score Vs Generated bias percentage

The behaviour of the polynomial regression equation was qualitatively similar to symbolic regression with the key difference: the trend of the score. The score calculated using the equation generated through polynomial regression displayed a consistent increase with the increasing levels of bias while the equation generated through symbolic regression displayed a gradual decline beyond a 40% bias threshold despite the increasing levels of bias. Despite the score increasing with the increasing levels of bias the score values could not quantify the degree of bias present in the dataset. To address the limitation min-max scaling was used to scale the score into a more interpretable range.

Based on the understanding of the score it was found, a positive score indicated the dataset to be biased towards male and negative score indicated the dataset to be biased towards females. Based on polarity, the min-max scaling was customized. If the polarity of the calculated score was positive, the maximum bound was derived by assigning the female associated feature values to a minimum and minimum bound was derived by assigning equal values to both the female and male associated

features. When the polarity of the calculated score was negative the process was reversed, and corresponding bounds were calculated. The normalization approach performed more consistently across the test datasets.

Through Evaluation of performance of both equations it was found that the equation generated through polynomial regression worked best for the intended metric.

2.4. Validation.

To validate the proposed bias metric, we employed the Word Error Rate (WER) — one of the most prominent and widely accepted evaluation metrics in speech recognition and bias detection in audio-based AI systems. WER offers insight into model-based bias, whereas our newly developed Bias Score is designed to capture data-inherent bias, independent of the model's architecture or training process.

For this validation, we used the Whisper-tiny model developed by OpenAI [30]. This model was chosen due to its multilingual support and lightweight nature, making it suitable for consistent testing across diverse datasets and languages. All datasets used for validation were passed through the Whisper-tiny model to obtain WER values for both male and female speakers separately.

A system is considered biased towards a specific gender when its WER is consistently lower for one gender compared to the other. For instance, if the WER is significantly higher for female speakers, the system is said to be male-biased, indicating that the model performs better on male audio data.

On the other hand, the Bias Score is calculated based on statistical audio features (e.g., pitch, energy, speech activity) extracted directly from the dataset, not the model output. A positive Bias Score indicates the dataset is biased towards males, while a negative score implies female bias.

The following table summarizes the WER values per gender, the bias interpretation based on WER, the corresponding Bias Score, and its interpretation. This comparison helps validate whether the Bias Score aligns with the bias suggested by WER:

Datasets	WER male	WER female	Bias defined by WER	Bias score	Definition of the score
LibriSpeech: Dev split	20.05	26.3	System : Male biased	0.650504	Dataset : Male Biased
LibriSpeech: Test-other split	16.0625	19.0588	System : Male biased	1.263191	Dataset : Male Biased
LibriSpeech: Test-clean split	22.45	24.75	System : Male biased	0.811159	Dataset : Male Biased
LibriSpeech Train split	31.5104	32.0797	System : Male biased	0.713163	Dataset : Male Biased
Multi-lingual LibriSpeech : Portuguese	1.148	1.2917	System : Male biased	-4.79006	Dataset : Female Biased
Multi-lingual LibriSpeech : Polish	1.222	1.2009	System : Female biased	6.279849	Dataset : Male Biased
Common voice : Hakha Chin Train split	1.17	1.2114	System : Male biased	1.259123	Dataset : Male Biased
Common voice : Hakha Chin other-split	1.1913	1.1497	System : Female biased	-5.24067	Dataset : Female Biased
Common voice : Hakha Chin validated split	1.1816	1.1599	System : Female biased	-0.23507	Dataset : Female Biased
Common voice : Chuvash test split	1.2571	1.2261	System : Female biased	-3.51048	Dataset : Female Biased
Common voice : Chuvash validated split	1.2433	1.2966	System : Male biased	2.512189	Dataset : Male Biased
Common voice : Sorani validated split	1.2121	1.4281	System : Male biased	8.25787	Dataset : Male Biased
Common voice : Sorani other split	1.1745	1.5312	System : Male biased	8.358887	Dataset : Male Biased
Common voice : Western Frisian other split	1.6055	1.3719	System : Female biased	-4.88802	Dataset : Female Biased
Common voice : Western Frisian validated split	2.7478	1.4275	System : Female biased	-6.17234	Dataset : Female Biased
Common voice : Indonesian validated split	1.5453	1.479	System : Female biased	-7.78833	Dataset : Female Biased
Common voice : Latgalian validated split	1.4751	1.4529	System : Female biased	-6.42433	Dataset : Female Biased
Common voice : Western Mari validated split	1.2638	1.2416	System : Female biased	-5.45872	Dataset : Female Biased
Common voice : Tartar validated split	1.1325	1.8452	System : Male biased	8.271134	Dataset : Male Biased
Common voice : Cantonese other split	1.3127	1.0833	System : Female biased	-5.7899	Dataset : Female Biased

Table 1 WER vs Bias Score

As seen in the table above, in most dataset splits, both the Bias Score and the WER-based interpretation agree on the dominant gender bias. The LibriSpeech splits consistently show a male bias, both in WER and Bias Score. The Sorani and Tartar splits have significantly higher Bias Scores (e.g., >8), which aligns with a system-level male bias indicated by higher female WERs. In a few cases, such as Multilingual LibriSpeech: Polish, there is a contradiction, where WER suggests a female-biased system, but the Bias Score shows male-biased data. These discrepancies may arise from model-specific error patterns or data complexity not captured directly by statistical features.

Such inconsistencies reinforce the idea that while WER is helpful, it is not purely reflective of dataset bias — it also captures model biases, vocabulary familiarity, and architecture sensitivity. In contrast, the Bias Score remains independent of the model and strictly analyzes the data-driven disparities, offering a complementary perspective for gender bias assessment in speech datasets.

3. RESULTS AND DISCUSSION.

3.1. Introducing The Tests Used For Performance Measurement.

I. Mean Squared Error (MSE).

MSE is a fundamental metric used to measure the average of the squared differences between predicted and actual values. A lower MSE value indicates better model performance.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- y_i = actual value
- $\hat{y}_i$ = predicted value
- n = number of data points

II. Root Mean Squared Error (RMSE).

RMSE is the square root of the MSE. It retains the same units as the target variable and provides an interpretable measure of average prediction error.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

III. Root Mean Squared Error (RMSE).

NMSE provides a scale-invariant version of the MSE by dividing it by the variance of the actual values. It helps compare model performance across datasets with different scales.

$$NMSE = \frac{MSE}{Var(y)}$$

- $VAR(y)$ = variance of the actual values.

IV. R-squared (R^2) – Coefficient of Determination

R^2 represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with values closer to 1 indicating a better fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - (y_i - \bar{y_i}))^2}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2}$$

- $\bar{y}_i$= mean of the actual values.

V. Mean Absolute Error (MAE)

MAE measures the average magnitude of errors in predictions without considering their direction. It is more robust to outliers than MSE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$$

VI. Pearson's Correlation Coefficient (r)

Pearson's r measures the linear correlation between actual and predicted values. Values range between -1 (perfect negative correlation) and +1 (perfect positive correlation), with 0 indicating no linear correlation.

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y_i} - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y_i} - \bar{\hat{y}})^2}}$$

VII. Spearman's Rank Correlation Coefficient (ρ).

Spearman's ρ is a non-parametric measure that assesses how well the relationship between two variables can be described using a monotonic function. It uses the ranks of values instead of raw values. Where, $\rho = +1$ is a perfect positive rank correlation, $\rho = -1$ is a perfect negative rank correlation and $\rho = 0$ is when no correlation is found in the ranks.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

- d_i= difference between the ranks of the actual and predicted values.

VIII. Kendall's Tau (τ)

Kendall's Tau is another non-parametric measure of ordinal association. It compares the number of concordant and discordant pairs. Like Spearman's ρ, the value of Kendall's τ also ranges from -1 to +1, with 0 indicating no ordinal relationship between the variables.

$$\tau = \frac{(C - D)}{\frac{1}{2} n(n - 1)}$$

- C = number of concordant pairs
- D = number of discordant pairs

3.2. Performance Measurements.

Dataset	MSE	RMSE	MAE	NMSE	r^2
LibriSpeech (LS)	0.07173	0.26784	0.1724	0.00209	0.9979
Multilingual LibriSpeech: Italian (IT)	0.02688	0.16395	0.1149	0.00078	0.9992
Multilingual LibriSpeech: Portuguese (PT)	1.23754	1.11245	0.4622	0.03622	0.9638
Multilingual LibriSpeech: Polish (PL)	0.35646	0.59704	0.3112	0.01043	0.9896
Common voice : Hakha Chin (CNH)	0.57049	0.75530	0.5480	0.01669	0.9833
Common Voice : Chuvash (CV)	1.01335	1.00665	0.7063	0.02965	0.9703
Common Voice : Welsh (W)	0.00097	0.03125	0.0254	2.85968	0.9998
Common Voice : Kurmanji (KMJ)	0.01545	0.12433	0.0990	0.00045	0.9995
TedLium (TDL)	0.26426	0.51406	0.3697	0.00773	0.9923
The AMI Corpus (AMI)	0.00503	0.07096	0.0517	0.00014	0.99985

Table 2 MSE, RMSE, NMSE, R-Squared, MAE

Dataset	Pearson's Correlation	Spearman's Rank Correlation	Kendall Tau Rank Correlation
LibriSpeech (LS)	0.9991	0.99951	0.99534
Multilingual LibriSpeech: Italian (IT)	0.9997	0.99983	0.99767
Multilingual LibriSpeech: Portuguese (PT)	0.9848	0.98654	0.95348
Multilingual LibriSpeech: Polish (PL)	0.9954	0.99485	0.97558
Common voice : Hakha Chin (CNH)	0.9972	0.99752	0.98255
Common Voice : Chuvash (CV)	0.9962	0.99760	0.98486
Common Voice : Welsh (W)	0.9993	0.999999	0.999999
Common Voice : Kurmanji (KMJ)	0.9999	0.99999	0.99999
TedLium (TDL)	0.9970	0.99756	0.98604
The AMI Corpus (AMI)	0.9999	0.99999	0.99999

Table 3 Pearson's Correlation, Spearman's Rank, Kendall Tau Rank

I. LibriSpeech.

LibriSpeech yielded outstanding performance across all metrics. The model produced a Mean Squared Error (MSE) of 0.07173 and an R-squared value of 0.9979, indicating that nearly all variance in the target variable was explained by the model. The correlation metrics reinforce this excellent fit: the Pearson's Correlation was 0.9991, Spearman's Rank Correlation was 0.99951, and Kendall's Tau was 0.99534. These high values suggest an extremely strong linear and monotonic relationship between actual and predicted scores.

II. Multilingual LibriSpeech: Italian.

For the Italian subset of the multilingual LibriSpeech, the model's performance was exemplary. With a very low MSE of 0.02688 and an R-squared of 0.9992, the predictions closely matched the true bias scores. Furthermore, the correlation coefficients were near perfect: Pearson's at 0.9997, Spearman's at 0.99983, and Kendall's Tau at 0.99767, showing a very strong agreement in both magnitude and order of values.

III. Multilingual LibriSpeech: Portuguese.

This subset had relatively higher errors, with an MSE of 1.23754 and a lower R-squared value of 0.9638, suggesting the model struggled to generalize to Portuguese. The correlation values with Pearson's at 0.9848, Spearman's at 0.98654, and Kendall's Tau at 0.95348, while still high, were lower compared to other datasets. This could indicate variations in language structure or acoustic features that were less well captured by the model.

IV. Multilingual LibriSpeech: Polish

The Polish dataset yielded solid results, with an MSE of 0.35646 and R-squared of 0.9896. The correlation metrics were also strong: Pearson's was 0.9954, Spearman's was 0.99485, and Kendall's Tau was 0.97558. These results indicate that the model performed reliably, maintaining strong predictive ordering and correlation despite moderate error values.

V. Common Voice: Hakha Chin

For Hakha Chin, the model achieved an MSE of 0.57049 and an R-squared of 0.9833, suggesting a good fit with slightly higher residual errors. The correlation metrics—Pearson's at 0.9972, Spearman's at 0.99752, and Kendall's Tau at 0.98255—show that the predicted values preserved both the strength and rank of relationships remarkably well, despite some variance in absolute errors.

VI. Common Voice: Chuvash

Chuvash data showed one of the weaker performances with an MSE of 1.01335 and R-squared of 0.9703. Nevertheless, the correlation values remained high: Pearson's at 0.9962, Spearman's at 0.99760, and Kendall's Tau at 0.98486. This suggests that

even though the model's predictions were less accurate in magnitude, the relative order and association between variables were still well preserved.

VII. Common Voice: Welsh.

Welsh data showed exceptional results with almost negligible error—MSE of 0.00097 and R-squared of 0.9998. The correlation metrics mirrored this perfection, with Pearson's Correlation at 0.9993, and both Spearman's and Kendall's Tau at 0.999999. These values indicate near-perfect linear and rank correlation, making Welsh one of the best-performing datasets in this study.

VIII. Common Voice: Kurmanji

The Kurmanji subset also performed extremely well. With an MSE of 0.01545 and R-squared of 0.9995, the model had minimal prediction error. Correlation values were similarly high—Pearson's at 0.9999, Spearman's at 0.99999, and Kendall's Tau at 0.99999—indicating an excellent fit across both value and ranking perspectives.

IX. TED-LIUM

TED-LIUM data resulted in a good model fit with an MSE of 0.26426 and R-squared of 0.9923. The correlation coefficients (Pearson's at 0.9970, Spearman's at 0.99756, and Kendall's Tau at 0.98604) suggest that the predictions were consistently aligned with the true values in both order and magnitude, despite slightly higher errors likely due to the spontaneous nature of TED talks.

X. AMI Corpus

The AMI Corpus yielded one of the most precise performances. The MSE was only 0.00503, and the R-squared was virtually perfect at 0.99985. Correlation values also reached maximum thresholds, with Pearson's at 0.9999, and both Spearman's and Kendall's Tau at 0.99999. This confirms that the model achieved nearly perfect alignment between predicted and actual bias scores in this dataset.

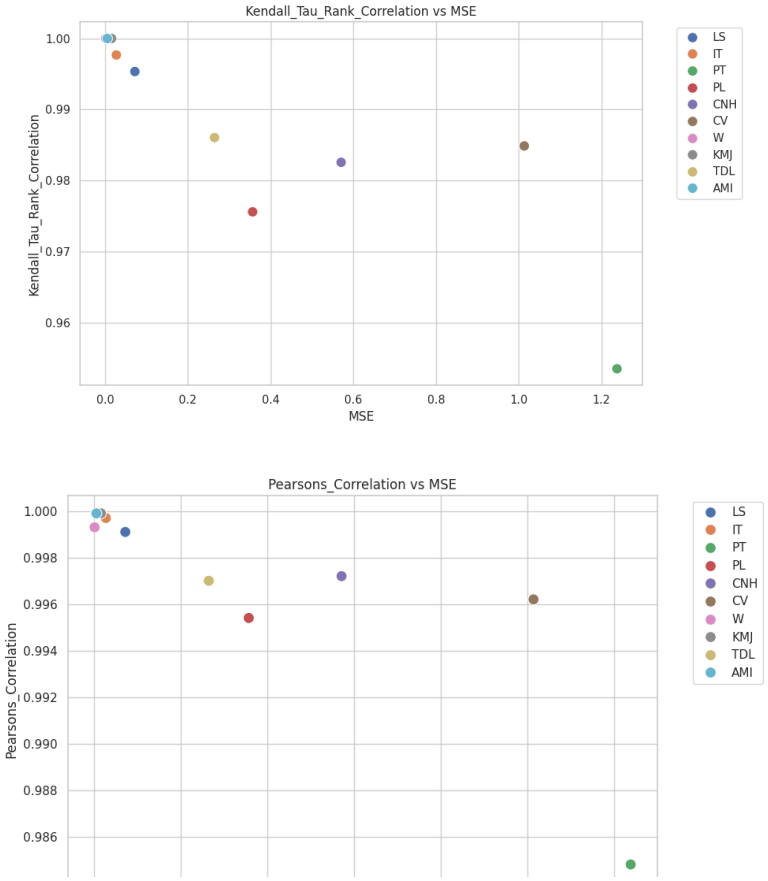

Figure 28 Kendall Tau Rank Vs MSE (top) , Pearson's Correlation Vs MSE (bottom)

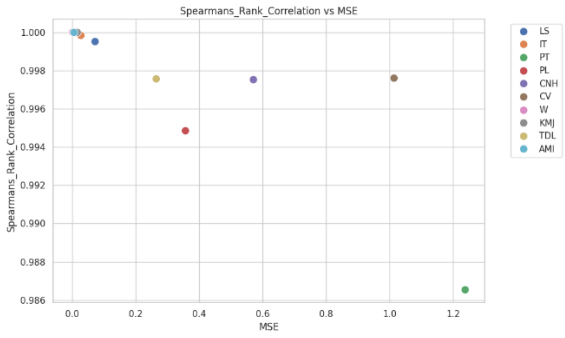

Figure 29 Spearman's Rank Vs MSE

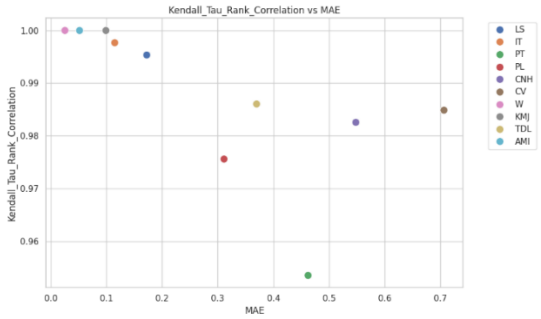

Figure 30 Kendall Tau Rank Vs MAE

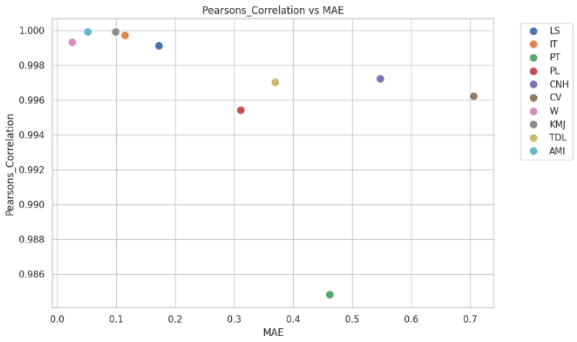

Figure 31 Pearson's Correlation Vs MAE

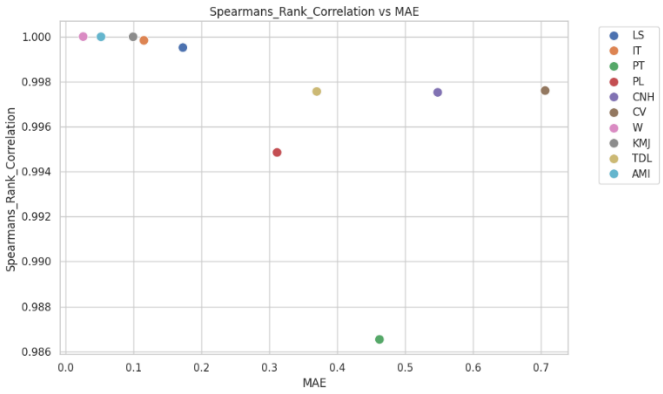

Figure 32 Spearman's Correlation Vs MAE

3.3. Discussion.

While the proposed bias detection metric introduces a novel and insightful approach to quantifying gender bias in audio datasets, it is important to acknowledge several limitations inherent in the current version of the metric. These limitations provide both a realistic boundary to its present capabilities and a roadmap for future enhancements and research directions.

The most significant limitation of the current metric is its exclusive focus on gender as the demographic dimension for bias analysis. While gender is a critical axis of evaluation especially in the context of fairness in AI systems, it is only one of many possible demographic attributes that can influence and skew dataset distributions and model behaviour. Important dimensions such as race, age, accent, socio-economic background, native language, and regional dialect are not captured in the current formulation of the metric. These factors, individually or in combination, can significantly affect the performance of speech-based AI systems and introduce systematic disparities that may go unnoticed if only gender is considered. For instance, the same dataset that appears balanced by gender might still demonstrate strong bias against older speakers, speakers from certain ethnic groups, or non-native speakers of a particular language. By not incorporating these dimensions, the metric potentially overlooks a wide array of intersecting and compounding biases, which could lead to incomplete or misleading interpretations of fairness in a dataset.

Another key limitation lies in the reliance on the availability of explicit gender metadata associated with the audio samples. The metric assumes that gender information is known and accurately labeled for each audio file. However, this is not always the case in real-world scenarios. Many publicly available or crowd-sourced speech datasets, such as parts of Mozilla Common Voice or TEDx corpora do not consistently provide speaker-level demographic information, or may offer gender labels based on self-identification without verification. In situations where such metadata is missing, ambiguous, or unreliable, the metric becomes either inapplicable or prone to error. While automated gender detection techniques exist, they introduce their own risks of inaccuracy and bias, particularly when dealing with

speakers whose voices deviate from typical male/female vocal patterns (e.g., children, elderly, or transgender individuals). As a result, the usability of the metric is currently confined to datasets where gender metadata is not only available.

The current implementation of the metric has been tested and validated primarily on single-speaker audio datasets. These datasets typically contain recordings where one individual speaks at a time, and the gender label is directly associated with that speaker. However, in datasets which contain meeting transcription, interview summarization, or conversational AI, audio files contain multi-speaker interactions, where multiple individuals participate in overlapping or turn-based speech. Analysing such datasets requires speaker diarization, a complex preprocessing step that segments the audio into speaker-specific portions and attributes each segment to an identified speaker. The proposed metric does not yet support or accommodate this level of complexity, and its effectiveness on multi-speaker datasets remains untested. Furthermore, errors introduced during diarization (e.g., speaker boundary errors, incorrect gender attribution) can compound the measurement inaccuracies of the bias metric.

while the proposed metric serves as a valuable starting point for understanding and quantifying gender bias in audio datasets, its current formulation has important limitations that constrain its broader applicability

4. CONCLUSION.

This study introduces a novel and robust metric designed to quantify the gender bias inherent within an audio-based dataset. The proposed metric places emphasis on utilizing raw audio features, ensuring that the analysis remains unaffected by any bias potentially introduced during the algorithmic processing or model-building phases. By targeting the dataset itself, rather than the outcomes or predictions made by machine learning models, the metric seeks to isolate and evaluate the bias that may already exist prior to any algorithmic intervention. This approach addresses a critical gap in current practices, which often overlook biases present at the dataset level, focusing instead on the model's performance without acknowledging the foundational issues that might contribute to skewed or unfair results.

Throughout the research, multiple methodologies were explored to develop a comprehensive and reliable metric. Various statistical techniques and machine learning models were considered to detect gender bias, and several iterations were tested to determine the most effective approach. Among the different strategies evaluated, the metric constructed using polynomial regression with L2 regularization emerged as the most promising. This specific configuration not only provided the most accurate results in terms of detecting bias but also offered a stable framework for quantifying bias in diverse datasets, demonstrating its scalability and robustness.

Additionally, the validation process involved comparing the performance of the newly proposed metric against established metrics such as Word Error Rate (WER), a traditional benchmark for assessing model-based bias. This validation process confirmed that the new metric aligns well with conventional methods of bias detection, while providing a significant advantage in that it isolates and quantifies bias at the dataset level, independent of any external model influences. The ability to distinguish between dataset-driven bias and model-induced bias is a key innovation of this study, as it enables a deeper understanding of the factors contributing to gender bias in audio datasets.

The metric introduced in this study represents a crucial first step toward building fairer, more equitable audio-based systems. By providing developers and researchers

with a reliable tool to identify and measure gender bias within their datasets, this metric empowers stakeholders to take proactive steps to mitigate such biases during the early stages of system development. This ability to detect and address bias at the dataset level opens up new possibilities for designing algorithms and models that are inherently fairer and more unbiased, leading to the creation of more inclusive systems that can better serve diverse populations.

As the field of audio-based machine learning continues to evolve, the development of this metric lays the groundwork for future research aimed at creating truly fair and unbiased systems. By recognizing and addressing the bias embedded in datasets, the work presented here contributes to the growing body of research focused on enhancing fairness in artificial intelligence. Ultimately, this metric is an essential tool for developers and researchers striving to create ethical AI systems, and it offers a pathway for future advancements in bias detection, correction, and mitigation across diverse domains.

In conclusion, the introduction of this metric offers a promising steppingstone for the identification and quantification of gender bias at the dataset level. It acts as a steppingstone for future efforts in developing fairer, more equitable audio-based systems, with the potential to significantly influence the development of AI technologies in a manner that prioritizes fairness, equity, and inclusivity.

5. 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] 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.

[10] 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.

[11] 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.

[12] s. Feng , O. Kudina, B. M. Halpern and O. Scharenborg, “Quantifying Bias in Automatic Speech Recognition,” *arXiv preprint arXiv:2103.15122,* 2021.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.