Investigating the Impact of Multilingual Pre-trained Speech Models on Gender Bias in ASR for Low Resource African Languages

Claytone Sikasote[1,2][0009−0004−1372−4662], Hussein Suleman[1][0000−0002−4196−1444], and Jan Buys[1][0000−0003−1994−5832]

[1] University of Cape Town, Cape Town, South Africa
skscla001@myuct.ac.za, hussein@cs.uct.ac.za, jan.buys@uct.ac.za
[2] University of Zambia, Lusaka, Zambia
claytone.sikasote@cs.unza.zm

Abstract. While fine-tuning transformer-based pre-trained speech models improves speech recognition for low resource languages, the approach increases the risk of speaker attribute bias in the resulting target language automatic speech recognition (ASR) systems. This work investigates gender bias in two state-of-the-art pre-trained speech models, MMS and Whisper, fine-tuned for ASR on three African languages: Bemba, Nyanja, and Swahili. We fine-tune models on gender-specific as well as gender-balanced datasets, and estimate and compare gender bias across different settings. Our results show varying degrees of gender bias in the fine-tuned models, even with gender-balanced fine-tuning, suggesting influence from pre-trained models. Inconsistencies in gender-specific fine-tuning further confirm the transfer of bias from pre-trained models. Additionally, an ablation study shows no relationship between training data size and gender bias.

Keywords: Gender Bias · Automatic Speech Recognition · Low Resource Languages · African languages.

1 Introduction

Fine-tuning transformer-based multilingual pre-trained speech models has become a dominant approach to the development of automatic speech recognition (ASR) systems for low-resource languages (LRLs) [5, 34, 33]. Typically, these models are pre-trained on large quantities of multilingual speech data (labeled or unlabeled) and later fine-tuned on a relatively small labeled target language dataset for a downstream speech task. While this approach has led to improved model performance for LRLs, it increases the risk of introducing speaker attribute-specific bias [12], which is the disparity in model performance between different speaker attribute subgroups, for example between male and female speakers.

ASR systems fine-tuned from pre-trained speech models have been reported to exhibit gender bias [4], performing better on male speech in some studies [16,

27] and female speech in others [12, 15, 26]. Recent studies have investigated pre-training data as a potential source of gender bias [29, 42]. While these studies offer insight into the impact of data composition in the pre-training data for bias mitigation, they do not investigate the impact of fine-tuning pre-trained speech models on gender bias in resulting ASR systems. In this work, we seek to address two questions: First, to what extent does fine-tuning pre-trained speech models impact gender bias in ASR systems fine-tuned on LRLs? Second, does the target language training data size affect gender bias in ASR systems fine-tuned on LRLs?

Our work considers a multilingual setup where access to the pre-training data is not available, and the evaluation focuses on gender bias in ASR systems fine-tuned on LRLs. Concretely, we investigate two state-of-the-art pre-trained models, Massively Multilingual Speech (MMS) and Whisper, evaluated on three Sub-Saharan African languages: Bemba (bem), Nyanja (nya), and Swahili (swa). All belong to the Niger-Congo Bantu family of languages. We leverage publicly available datasets with sufficient data labelled by gender. To our knowledge, this is the first work that comprehensively investigates the impact of fine-tuning pre-trained speech models on gender bias for LRLs spoken in Sub-Saharan Africa.

We evaluate pre-trained models after fine-tuning, estimating and comparing gender bias after gender-specific fine-tuning and after fine-tuning on datasets with different degrees of gender balance. Moreover, we investigate whether there is a relationship between training data size and gender. Our results show that fine-tuning on a gender-balanced dataset does not necessarily mitigate gender bias in resulting ASR systems for LRLs, suggesting potential influence from the pre-trained model. In addition, our ablation study reveals that, while speech recognition performance improves with an increase in training data size, there is no relationship between training data and gender bias, suggesting dependence on the dataset of the target language.

The rest of the paper is structured as follows. In Section 2, we provide related work. Section 3 outlines our methodology. Section 4 presents and discusses the experimental results. Finally, in Section 5 we draw conclusions and suggest possible directions for future work.

2 Related Work

Bias in artificial intelligence (AI) systems, in general, remains a challenge for developing inclusive technology systems [28]. Research shows that AI systems in various domains, such as computer vision [11], natural language processing [28], and speech processing [31], exhibit biases that, if not mitigated, have the potential to perpetuate discrimination against certain groups of users in society [7]. In this section, we outline related works that investigate different forms of speaker attribute bias in ASR systems, including in pre-trained speech models.

2.1 Bias in ASR Systems

Several studies across a range of languages have investigated ASR systems for different forms of speaker attribute bias. With respect to gender bias, studies show that ASR systems often recognize speech of one gender more accurately than another: some report a higher recognition accuracy for female speakers [1, 19, 24, 35], while others observe a better performance for male speech [16, 39]. In other cases, no noticeable difference in speech recognition accuracy is observed between male and female speakers [40]. In addition to gender bias, ASR systems have also been found to exhibit bias based on race [24], native language [12], age [13, 14], dialects [19, 39], and nationality [23]. These studies investigate and quantify bias in ASR systems, primarily, for high-resource languages, such as English [1, 19, 24], French [1, 16], Portuguese [26, 27], Dutch [12, 21, 43, 44], Mandarin [12], Spanish [10, 30], and Arabic [35]. For LRLs, where much of the research efforts are focused on creating training data and improving model accuracy [36–38], minimal attention is paid to investigating biases models exhibit and how to mitigate them. In this study, we investigate gender bias in low resource African languages.

2.2 Multilingual Speech Models

With regard to pre-trained speech models, studies have investigated multilingual models, which have improved the speech recognition accuracy for most LRLs [26, 15]. Kulkarni et al. [26] investigate bias in the Massively Multilingual Model (MMS) and Whisper, and demonstrate that fine-tuning these models for Mexican Spanish results in models exhibiting gender bias against female speech, and with no noticeable bias for age, accent, and skin tone color. Similarly, Fuckner et al. [15] investigate Whisper and XLS-R [5] for Dutch speakers, and find performance disparities for non-native, children, and elderly speakers. Gody and Harwath [17] investigate HuBERT [22] for topic diversity, number of speakers, and speaker gender. Beyond these off-the-shelf foundational models, there are multilingual pre-trained speech models that have also been developed specifically for African languages. Caubriere and Gauthier [9] pre-trained the first self-supervised learning multilingual speech model based on 60K unlabeled speech data covering 21 Sub-Saharan African languages. The model is a 95 million parameter model based on the HuBERT [22] base architecture. Most recently, Alabi et al. [2] presented AfriHuBERT, which is an extension of the 95 million parameter mHuBERT-147 [8] model pre-trained on 10K hours of speech data covering over 1200 African languages from diverse sources. Although these models are yet to be investigated for different forms of speaker attribute bias, our work investigates the Whisper and MMS models, which are larger in size, and examines the extent to which fine-tuning these models on low resource African languages impacts gender bias.

3 Methodology

In this section, we outline our methodology to address our research questions. We first describe the datasets we use for our experiments in Section 3.1. In Section 3.2, we describe the speech models we investigate. We provide our fine-tuning and evaluation setup in Section 3.3.

3.1 Datasets

We use publicly available datasets for a selection of Niger-Congo Bantu languages that have a sufficient amount of data with gender annotations in their metadata. [3] The dataset statistics in Table 1 show that the gender distribution in most of these datasets is severely imbalanced. We use the following datasets:
BembaSpeech (BS) is a monolingual ASR dataset for Bemba, comprising more than 20 hours of read speech recorded by 17 Bemba speakers: 9 male and 8 female, using text sourced from public literature [36].
Bemba Image-Grounded Conversations (BIG-C) is a multimodal and multi-purpose dataset that includes more than 180 hours of conversational speech in Bemba recorded by Bemba speakers based on images [37]. The images used to create speech in this dataset are obtained from the publicly available image dataset: Flickr30K [20].
Zambezi Voice (ZV) is a multilingual speech corpus covering 4 languages: Bemba, Nyanja, Lozi, and Tonga, which are spoken principally in Zambia [38]. The dataset provides both transcribed and untranscribed speech data. We use the transcribed read speech for the Nyanja set in our study. The Lozi and Tonga sets do not have gender annotation in the metadata files of the audio files.
Common Voice (CV) is a multilingual speech corpus that comprises read speech in more than 60 languages. We use the Swahili data in the November 2024 (v20) release [3]. A substantial proportion of the data does not have gender in its metadata, but due to the size of the dataset, we are able to only use the gender-annotated subset.

3.2 Pre-trained Speech Models

We investigate two state-of-the-art multilingual pre-trained speech models:
MMS is a Wav2Vec2.0 [6] based multilingual speech model pre-trained on approximately half a million hours of audio comprising more than 1100 languages [32]. The models support a range of tasks, including speech recognition, speech-to-text translation, language identification, and speech synthesis. MMS model variants are based on two model parameter sizes: 317 million and 965 million parameter base models. In our study, we use the checkpoint of a 965 million multilingual MMS model that is fine-tuned from the 965 million base model.

[3] We investigate gender bias as a binary construct (i.e., male and female) based on practical constraints, relying on annotations in the metadata of the considered datasets.

Whisper is a family of pre-trained models trained on more than 680,000 hours of multilingual and multi-task data collected from the web in a weakly supervised approach in 98 languages [34]. Whisper models come in several sizes: Tiny (39 million), Base (74 million), Small (244 million), Medium (769 million), Large, Large-v2, and Large-v3. All large models have 1550 million parameters. In this work, we use the medium model.

Table 1. Dataset sizes (in hours) of the gender split in the original datasets and in our gender-balanced subsets. The original datasets include BembaSpeech (BS), Bemba Image Grounded Conversations (BIGC), Zambezi Voice (ZV), and Common Voice (CV). *Male* and *Female* data split sizes are given, along with a *Baseline* subsampled according to the original distribution and gender-balanced subsets *Balanced* and *Combined*. The *Balanced* dataset has the same size as the *Male* and *Female* sets, while the *Combined* is a combination of the two gender-specific datasets (Male and Female).

	Original			**Gender-Balanced**			
Datasets	**Male**	**Female**	**Baseline**	**Male**	**Female**	**Balanced**	**Combined**
Bemba-BS							
train	13.0	7.1	6.9	6.9	6.8	6.9	13.7
valid	1.7	0.8	0.9	0.9	0.9	0.9	1.7
test	1.4	0.6	-	0.9	0.8	-	1.7
Bemba-BIGC							
train	92.5	74.6	5.1	5.0	5.0	5.0	10.0
valid	3.3	2.6	0.9	0.9	0.9	0.9	1.8
test	3.2	2.6	-	0.9	0.9	-	1.8
Nyanja-ZV							
train	18.5	2.3	4.2	4.4	4.3	4.3	8.7
valid	0.3	1.9	0.5	0.5	0.5	0.5	1.0
test	0.2	1.2	-	0.5	0.5	-	1.0
Swahili-CV							
train	19.5	24.1	5.1	5.1	5.2	5.1	10.3
valid	4.6	6.2	0.9	0.9	0.9	0.9	1.8
test	4.7	5.6	-	0.9	0.9	-	1.8

3.3 Experimental Setup

We study the impact of fine-tuning pre-trained speech models on gender bias in ASR systems for LRLs through settings that control the training data and hyperparameters to mitigate confounding factors such as data bias.

Evaluations We evaluate the performance of our models using the standard ASR metric: Word Error Rate (WER). For *Bias*, we adopt the definition of Feng et al. [12] as the difference in WER between the different speaker groups

for each attribute-specific dimension. We use the Kruskal-Wallis H-test [25] to determine the statistical significance of the difference in the distributions of the model performance between the gender groups.

Curating and Creating Training Datasets Given the disparity in the amount of speech data for the gender subgroups in the splits of the target datasets, we create artificial datasets with varying proportional representations of gender subgroup speech for our experiments. We randomly downsample the overrepresented group in the original dataset of the target language to create five datasets: *gender-specific* datasets, *Male* and *Female*, comprising of gender-specific speech data for training gender-specific models; a *Balanced* dataset of the same size as the gender-specific datasets, comprising of an equal representation of male and female speech; a *Baseline* dataset having the same proportional representation of speech by gender as the original dataset; and a *Combined* dataset (twice the size of the Balanced dataset), a combination of the two gender-specific datasets (Male and Female).

Given the size of the datasets and the small number of speakers in our setting, we carefully split the datasets into training, validation, and test sets based on the unique speaker IDs in the metadata file of the original datasets, ensuring that no speaker overlaps between the sets. We do this for *Bemba-BS* (derived from BembaSpeech) and Nyanja (from ZV). *Bemba-BIGC* (derived from BIG-C) and Swahili have sufficient data to create splits that are subsets of the original training/validation/test splits. Additionally, all datasets derived from BIG-C are balanced with an equal proportion of native and non-native speech to mitigate any potential bias due to native speaking style from becoming a confounding factor in our investigation. We preprocess the transcriptions by normalizing the text, removing the punctuation, and lower-casing the characters. Audio files are converted from MP3 to WAV format with a sample rate of 16Khz. Table 1 provides details of the original and the resulting data sets. We evaluate using the test sets from two gender-specific datasets and the combined dataset.

Fine-tuning Pre-trained Multilingual Models We use the 965 million parameter fine-tuned ASR MMS model[4] and the (medium) 769 million parameter Whisper model[5] for our fine-tuning experiments. We train gender-specific models using gender-specific datasets described in Section 3.3 and compare their performance with the models trained on balanced and combined sets (twice the size of gender-specific sets). All models are fine-tuned using the Hugging Face Transformer library [41] with the Connectionist Temporal Classification (CTC) criterion [18]. We fine-tune 3 different models with different training seeds on each dataset to ensure that we obtain reliable bias estimates, particularly given the low-resource nature of the datasets. Except for learning rates and batch size, we inherit the default configurations from the library for other parameter settings. We experiment with different learning rates (1e-4, 7e-4, 3e-4, 1e-5,

[4] https://huggingface.co/facebook/mms-1b-all
[5] https://huggingface.co/openai/whisper-medium

5e-5) and batch sizes (2, 4, 8). All models are trained for 30 epochs with an early-stopping set to 3. MMS models are trained with a learning rate of 3e-4, and Whisper is trained using 1e-5. For batch size, we use 8 for MMS and 2 for Whisper models. All models are trained on an A100 GPU.

Effect of Fine-tuning Data Size on Gender Bias To address our second question, we conduct an ablation study using Bemba and Swahili to investigate whether there is a correlation between fine-tuning data size and gender bias in fine-tuned ASR models. We create gender-balanced training sets with 5 to 30 hours of speech (with a 5-hour interval), following the approach described in Section 3.3. We use subsets of the two largest training datasets, Bemba BIG-C and Swahili CV. We fine-tune our target pre-trained models on these datasets using the same setup and configurations as above.

4 Results

4.1 Fine-tuning multilingual pre-trained speech models

Tables 2 and 3 present our results for the MMS and Whisper models, respectively. We report the *mean* along with the *standard deviation (STD)* of the WERs of the 3 different models fine-tuned and evaluated on each dataset. The names *Baseline*, *Male*, *Female*, *Balanced*, and *Combined* appended to the model names denote the fine-tuning datasets. We also conduct zero-shot evaluations of the models, i.e. without any fine-tuning. However, we report zero-shot results only for MMS and not for Whisper, since the Whisper-*ZeroShot* WERs were found to be higher than 100% in all cases and therefore we do not consider it viable to use these results in the gender bias analysis. *Bias* denotes the difference between male and female test set WERs, representing the degree of gender bias; a negative bias value implies that male speech was recognized more accurately than female speech.

Gender bias: We observe that both the MMS and Whisper models recognize female speech more accurately than male speech in Bemba-BS, Nyanja, and Swahili, while male speech is recognized better on the Bemba-BIGC dataset. This shows that the fine-tuned ASR models exhibit gender bias and that whether male or female speech are recognized more accurately depends on the language and dataset. In some cases, the bias is not significant; however, surprisingly, models with gender-specific training data sometimes exhibit lower bias than those trained on balanced data. Models trained on the two Bemba datasets exhibit gender bias in opposite directions, with Bemba-BS models consistently recognizing female speech more accurately and Bemba-BIGC models favouring male speech. We attribute this to the difference in speech type between the two training datasets: Bemba-BS comprises read speech, while Bemba-BIGC comprises conversational speech. These results suggest that the speaking styles of the speakers in the dataset may not only affect performance [12] but can potentially affect

which gender subgroup speech is favored by the ASR systems. We also observe that both models (MMS and Whisper) fine-tuned on gender-balanced datasets (*Balanced* and *Combined*) exhibit gender bias across most of the languages, with the exception of Swahili and Bemba-BIGC. When evaluating on gender-specific test sets, models fine-tuned on gender-specific datasets have the best WER or perform similarly to the best-performing model on the test set of the matching gender. There are a few cases though where the WER of models trained on one gender is lower when evaluating on the opposite gender test set than on the same gender test set. These results point to the pre-trained models as the likely source of influence for the gender bias in the fine-tuned models. The results of the baseline model further confirm this assertion. In some cases, the gender subset that is recognized more accurately by the baselines is under-represented in the training dataset. For example, female speech is recognized more accurately than male speech in Bemba-BS, even though it is very underrepresented in the original training dataset. This suggests that balancing gender speech in training datasets does not necessarily mitigate gender bias in fine-tuned ASR models for LRLs.

Lastly, in comparison to the results obtained on other datasets, the Bemba-BS dataset exhibits substantially higher bias estimates for both the MMS and Whisper models. Although female speech was recognized more accurately than male speech, a comprehensive error analysis is required to identify the types of errors produced by the models and to investigate their underlying causes. However, one possible explanation for the higher bias estimates is that the sentences used to generate the male test speech may have been more difficult than those used for the female test speech. This observation highlights the importance of developing bias evaluation datasets in which speech from each subgroup is recorded using the same set of sentences to enable fair and consistent comparisons.

Model performance: For both MMS and Whisper, we observe that models fine-tuned on mixed-gender datasets (*Baseline*, *Balanced*, and *Combined*) perform similarly or better than gender-specific models on the overall test set. We attribute this to the gender diversity in the balanced training datasets. In some cases, MMS baseline models (fine-tuned on *Baseline*) outperform models fine-tuned on both gender-specific (*Male* and *Female*) and balanced (*Balanced* and *Combined*) datasets. We do not observe this with Whisper models. However, in both MMS and Whisper models, models based on gender-balanced datasets (Balanced and Combined) perform similarly or better than those fine-tuned on biased datasets (*Baseline*, *Male*, and *Female*). This points to improved speech diversity in the gender-balanced speech data. The results suggest that one does not necessarily need to balance the training data by gender to obtain a good WER score. Models fine-tuned on the combined datasets perform similarly or better than models based on the balanced datasets, likely due to the increased training data size as the combined dataset is twice the size of the balanced dataset. This observation is similar to the findings in Meng et al. [29].

Table 2. Mean and standard deviation (in parentheses) of WERs (%) of MMS models on the test sets of the *Male*, *Female*, and *Combined* data splits. The results are averaged over 3 runs. The lowest WER is underlined for each column, and the lowest absolute gender bias (Bias) is in bold. Negative *Bias* values mean male speech was recognized more accurately than female speech.

Model	**Male**	**Female**	**Combined**	**Bias**
Bemba-BS				
MMS-*ZeroShot*	81.87 (0.00)	51.08 (0.00)	66.10 (0.00)	30.79 (0.00)
MMS-Baseline	56.20 (0.81)	34.29 (1.72)	44.99 (1.01)	21.91 (1.80)
MMS-Male	55.76 (1.75)	33.61 (1.12)	44.42 (1.36)	22.15 (1.09)
MMS-Female	57.26 (1.05)	35.06 (0.75)	45.89 (0.90)	22.20 (1.09)
MMS-Balanced	54.08 (0.58)	32.84 (0.45)	43.21 (0.47)	**21.24 (0.42)**
MMS-Combined	57.51 (3.74)	35.90 (4.17)	46.44 (3.96)	21.61 (0.45)
Bemba-BIGC				
MMS-*ZeroShot*	81.71 (0.00)	84.36 (0.00)	83.05 (0.00)	-2.65 (0.00)
MMS-Baseline	49.44 (4.55)	55.36 (2.21)	51.76 (4.15)	-5.92 (2.93)
MMS-Male	45.97 (1.52)	54.03 (1.51)	50.05 (1.51)	-5.37 (4.65)
MMS-Female	52.03 (2.51)	53.21 (1.68)	52.63 (2.09)	**-0.79 (0.90)**
MMS-Balanced	54.07 (7.40)	57.78 (6.14)	55.97 (6.78)	-2.47 (2.31)
MMS-Combined	51.49 (2.63)	55.24 (1.90)	53.40 (2.26)	-3.75 (0.79)
Nyanja-ZV				
MMS-*ZeroShot*	65.21 (0.00)	53.36 (0.00)	62.65 (0.00)	11.85 (0.00)
MMS-Baseline	34.02 (0.85)	29.47 (0.28)	33.04 (0.73)	4.55 (0.58)
MMS-Male	39.27 (1.00)	30.91 (0.78)	37.47 (0.88)	8.36 (0.95)
MMS-Female	34.42 (2.66)	30.58 (0.78)	33.59 (2.36)	3.85 (1.77)
MMS-Balanced	35.38 (2.73)	31.37 (0.42)	34.51 (2.16)	4.01 (2.69)
MMS-Combined	31.89 (1.53)	28.33 (0.58)	31.12 (1.23)	**3.56 (1.51)**
Swahili-CV				
MMS-*ZeroShot*	25.19 (0.00)	21.78 (0.00)	23.44 (0.00)	3.41 (0.00)
MMS-Baseline	26.44 (2.11)	24.89 (2.25)	23.12 (1.92)	3.05 (0.57)
MMS-Male	27.54 (3.41)	24.89 (3.75)	26.19 (3.58)	2.65 (0.35)
MMS-Female	24.86 (1.58)	21.46 (2.25)	23.12 (1.92)	3.40 (0.71)
MMS-Balanced	23.57 (0.16)	20.75 (0.09)	22.12 (0.07)	2.82 (0.20)
MMS-Combined	23.63 (0.12)	20.36 (0.32)	21.96 (0.20)	**2.59 (0.02)**

4.2 Effect of training data size on gender bias

Figure 1 shows a plot of training data size (hours) against performance (WER) of models fine-tuned on different amounts of gender-balanced training data and evaluated on the gender-specific tests (*Male* and *Female*). The aim of this ablation is to determine whether there is a relationship between training data size and gender bias, based on Bemba and Swahili. While we observe a reduction in WER as the amount of training data increases, we do not observe any systematic change in gender bias as the training data size increases. We conjecture that this shows that fine-tuning on an equal proportion of male and female speech

Table 3. Mean and standard deviation (in parentheses) of WERs (%) of Whisper models on the test sets of the *Male*, *Female*, and *Combined* data splits. The results are averaged over 3 runs. The lowest WER is underlined for each column, and the lowest absolute gender bias (Bias) is in bold. Negative *Bias* values mean male speech was recognized more accurately than female speech.

Model	**Male**	**Female**	**Combined**	**Bias**
Bemba-BS				
Whisper-*ZeroShot*	-	-	-	-
Whisper-*Baseline*	62.92 (0.67)	37.25 (0.34)	49.76 (0.42)	25.67 (0.63)
Whisper-Male	64.60 (1.07)	38.04 (1.34)	50.99 (0.18)	26.56 (1.56)
Whisper-Female	71.28 (1.16)	45.18 (1.75)	57.90 (1.39)	26.10 (1.11)
Whisper-Balanced	63.31 (0.27)	38.73 (0.58)	50.53 (0.18)	**24.58 (0.31)**
Whisper-Combined	59.98 (1.14)	34.36 (0.42)	46.85 (0.42)	25.62 (1.49)
Bemba-BIGC				
Whisper-*ZeroShot*	-	-	-	-
Whisper-*Baseline*	49.45 (0.10)	55.74 (2.14)	52.64 (1.03)	-4.19 (3.96)
Whisper-Male	56.31 (4.66)	59.88 (4.55)	58.12 (0.01)	-2.38 (6.83)
Whisper-Female	56.81 (2.83)	56.62 (4.24)	56.71 (3.55)	**0.13 (1.01)**
Whisper-Balanced	53.17 (2.72)	59.04 (6.55)	56.15 (1.99)	-3.91 (7.37)
Whisper-Combined	47.34 (1.63)	51.14 (2.70)	49.26 (2.18)	-2.54 (1.23)
Nyanja-ZV				
Whisper-*ZeroShot*	-	-	-	-
Whisper-*Baseline*	34.43 (0.59)	29.67 (0.06)	33.41 (0.47)	3.18 (2.78)
Whisper-Male	39.76 (0.68)	31.28 (0.65)	37.93 (0.40)	5.65 (4.99)
Whisper-Female	35.67 (2.21)	31.46 (0.13)	34.76 (1.70)	**2.81 (1.94)**
Whisper-Balanced	36.17 (3.05)	29.30 (0.45)	32.50 (1.76)	4.58 (4.67)
Whisper-Combined	32.77 (0.14)	28.43 (0.78)	31.52 (0.38)	2.90 (1.59)
Swahili-CV				
Whisper-*ZeroShot*	-	-	-	-
Whisper-*Baseline*	41.79 (9.33)	35.19 (4.78)	38.41 (7.00)	6.60 (4.56)
Whisper-Male	45.47 (2.28)	37.54 (0.47)	41.41 (0.88)	5.28 (4.97)
Whisper-Female	43.13 (10.13)	37.27 (0.03)	40.13 (4.93)	3.90 (7.94)
Whisper-Balanced	37.01 (1.00)	33.72 (1.02)	35.32 (0.95)	3.29 (0.63)
Whisper-Combined	32.25 (0.35)	31.64 (0.19)	31.94 (0.08)	**0.41 (0.52)**

data neither adds gender bias to the model nor reduces pre-existing bias consistently, therefore again pointing to the pre-trained model as the source of gender bias. However, the results do not indicate any relationship between training data size and gender bias. This confirms that attribute-specific bias, such as gender bias, remains a complex issue in ASR systems that is language- and dataset-dependent, regardless of which gender is more accurately recognized, in line with the findings in Attanasio et al. [4].

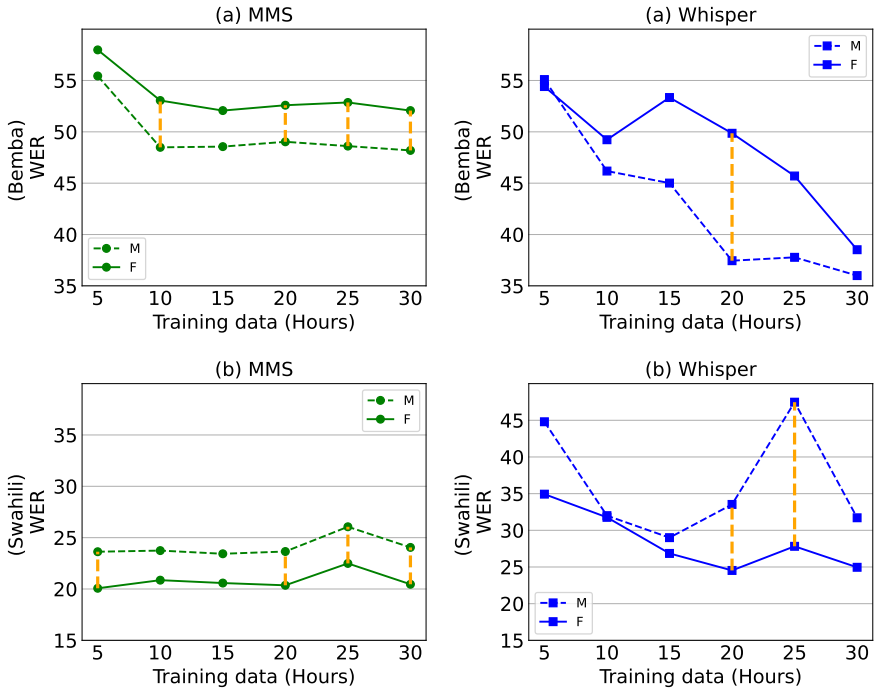

Fig. 1. Plot of training data size (hours) and model performance (WER) on gender subgroup test sets (*Male* and *Female*) of models fine-tuned on different degrees of training data. MMS models are in green while Whisper models are in blue. Solid lines represent model performance on the male test set, while dashed lines represent performance on the female test set for each respective model. Vertical orange dashed lines depict statistically significant disparities ($P < 0.05$).

5 Conclusion

We investigated the impact of fine-tuning multilingual pre-trained speech models on gender bias in ASR systems for LRLs, and whether there is a potential relationship between the size of the training data and gender bias. We find that fine-tuning Whisper and MMS models on gender-balanced datasets yields varying results across languages, with ASR systems exhibiting significant gender bias on gender-balanced datasets, suggesting potential bias propagation from the pre-trained models. We do not observe any correlation between training data size and gender bias, suggesting that gender bias is language- and dataset-dependent. Beyond collecting more speech data and improving speech diversity in training data, future work could explore developing fine-tuning strategies that aim to improve model performance and mitigate bias transfer from pre-trained speech models in low-resource settings.

Acknowledgments. This work was financially supported by the Hasso Plattner Institute (HPI) for Digital Engineering, through the HPI Research School at the University of Cape Town. This work is based on research supported in part by the National Research Foundation of South Africa (Grant Number 151601). Computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing (HPC) team: https://ucthpc.uct.ac.za/

References

1. Adda-Decker, M., Lamel, L.: Do speech recognizers prefer female speakers? In: Interspeech 2005. pp. 2205–2208 (2005). https://doi.org/10.21437/Interspeech.2005-699
2. Alabi, J.O., Liu, X., Klakow, D., Yamagishi, J.: AfriHuBERT: A self-supervised speech representation model for African languages. In: Interspeech 2025. pp. 4023–4027 (2025). https://doi.org/10.21437/Interspeech.2025-1437
3. Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., Weber, G.: Common voice: A massively-multilingual speech corpus. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 4218–4222. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.520
4. Attanasio, G., Savoldi, B., Fucci, D., Hovy, D.: Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 21318–21340. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.1188, https://aclanthology.org/2024.emnlp-main.1188/
5. Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., Auli, M.: Xls-r: Self-supervised cross-lingual speech representation learning at scale. In: Interspeech 2022. pp. 2278–2282 (2022). https://doi.org/10.21437/Interspeech.2022-143
6. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 12449–12460. Curran Associates, Inc. (2020)
7. Bender, E.M., Friedman, B.: Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics **6**, 587–604 (2018)
8. Boito, M.Z., Iyer, V., Lagos, N., Besacier, L., Calapodescu, I.: mHuBERT-147: A Compact Multilingual HuBERT Model. In: Interspeech 2024 (2024)
9. Caubrière, A., Gauthier, E.: Africa-centric self-supervised pre-training for multilingual speech representation in a sub-saharan context (2024), https://arxiv.org/abs/2404.02000
10. Chizhikova, A., Billinghurst, H., Elizabeth, M., Hossain, S., Kulkarni, A., Guibon, G., Couceiro, M.: Factorizing Gender Bias in Automatic Speech Recognition for Mexican Spanish (Sep 2024), https://hal.science/hal-04607587, working paper or preprint

11. Dehdashtian, S., He, R., Li, Y., Balakrishnan, G., Vasconcelos, N., Ordonez, V., Boddeti, V.N.: Fairness and bias mitigation in computer vision: A survey (2024), https://arxiv.org/abs/2408.02464
12. Feng, S., Halpern, B.M., Kudina, O., Scharenborg, O.: Towards inclusive automatic speech recognition. Computer Speech & Language **84**, 101567 (2024). https://doi.org/https://doi.org/10.1016/j.csl.2023.101567, https://www.sciencedirect.com/science/article/pii/S0885230823000864
13. Fenu, G., Marras, M., Medda, G., Meloni, G.: Fair voice biometrics: Impact of demographic imbalance on group fairness in speaker recognition. In: Interspeech 2021. pp. 1892–1896 (2021). https://doi.org/10.21437/Interspeech.2021-1857
14. Fenu, G., Medda, G., Marras, M., Meloni, G.: Improving fairness in speaker recognition. In: Proceedings of the 2020 European Symposium on Software Engineering. p. 129–136. ESSE 2020, ACM (Nov 2020). https://doi.org/10.1145/3393822.3432325, http://dx.doi.org/10.1145/3393822.3432325
15. Fuckner, M., Horsman, S., Wiggers, P., Janssen, I.: Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for dutch speakers. In: 2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD). pp. 146–151 (2023). https://doi.org/10.1109/SpeD59241.2023.10314895
16. Garnerin, M., Rossato, S., Besacier, L.: Gender representation in french broadcast corpora and its impact on asr performance. In: Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery. p. 3–9. AI4TV '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3347449.3357480, https://doi.org/10.1145/3347449.3357480
17. Gody, R., Harwath, D.: Unsupervised fine-tuning data selection for asr using self-supervised speech models. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10095264
18. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: ACM International Conference Proceeding Series (2006). https://doi.org/10.1145/1143844.1143891
19. Harris, C., Mgbahurike, C., Kumar, N., Yang, D.: Modeling gender and dialect bias in automatic speech recognition. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 15166–15184. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024), https://aclanthology.org/2024.findings-emnlp.890
20. Harwath, D., Glass, J.: Deep multimodal semantic embeddings for speech and images. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). pp. 237–244 (2015). https://doi.org/10.1109/ASRU.2015.7404800
21. Herygers, A., Verkhodanova, V., Coler, M., Scharenborg, O., Georges, M.: Bias in flemish automatic speech recognition. In: Draxler, C. (ed.) Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2023. pp. 158–165. TUDpress, Dresden (Mar 2023)
22. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3451–3460 (2021). https://doi.org/10.1109/TASLP.2021.3122291
23. Hutiri, W.T., Ding, A.Y.: Bias in automated speaker recognition. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and

Transparency. p. 230–247. FAccT '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3531146.3533089, https://doi.org/10.1145/3531146.3533089

24. Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J.R., Jurafsky, D., Goel, S.: Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences **117**(14), 7684–7689 (2020). https://doi.org/10.1073/pnas.1915768117, https://www.pnas.org/doi/abs/10.1073/pnas.1915768117
25. Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association **47**(260), 583–621 (1952), http://www.jstor.org/stable/2280779
26. Kulkarni, A., Kulkarni, A., Couceiro, M., Trancoso, I.: Unveiling biases while embracing sustainability: Assessing the dual challenges of automatic speech recognition systems. In: Interspeech 2024. pp. 4628–4632 (2024). https://doi.org/10.21437/Interspeech.2024-2494
27. Kulkarni, A., Tokareva, A., Qureshi, R., Couceiro, M.: The balancing act: Unmasking and alleviating ASR biases in Portuguese. In: Chakravarthi, B.R., B, B., Buitelaar, P., Durairaj, T., Kovács, G., García Cumbreras, M.Á. (eds.) Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion. pp. 31–40. Association for Computational Linguistics, St. Julian's, Malta (Mar 2024), https://aclanthology.org/2024.ltedi-1.4
28. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning (2022), https://arxiv.org/abs/1908.09635
29. Meng, Y., Chou, Y.H., Liu, A.T., Lee, H.y.: Don't speak too fast: The impact of data bias on self-supervised speech models. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3258–3262 (2022). https://doi.org/10.1109/ICASSP43922.2022.9747897
30. Nacimiento-García, E., Díaz-Kaas-Nielsen, H.S., González-González, C.S.: Gender and accent biases in ai-based tools for spanish: A comparative study between alexa and whisper. Applied Sciences **14**(11) (2024). https://doi.org/10.3390/app14114734, https://www.mdpi.com/2076-3417/14/11/4734
31. Ngueajio, M.K., Washington, G.: Hey asr system! why aren't you more inclusive? automatic speech recognition systems' bias and proposed bias mitigation techniques. a literature review. In: HCI International 2022 – Late Breaking Papers: Interacting with EXtended Reality and Artificial Intelligence: 24th International Conference on Human-Computer Interaction, HCII 2022, Virtual Event, June 26 – July 1, 2022, Proceedings. p. 421–440. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-21707-4_30
32. Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.N., Conneau, A., Auli, M.: Scaling speech technology to 1,000+ languages. J. Mach. Learn. Res. **25**(1) (Jan 2024)
33. Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., Collobert, R.: Mls: A large-scale multilingual dataset for speech research. In: Interspeech 2020. pp. 2757–2761 (2020). https://doi.org/10.21437/Interspeech.2020-2826
34. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)

35. Sawalha, M., Shariah, M.A.: The effects of speakers' gender, age, and region on overall performance of arabic automatic speech recognition systems using the phonetically rich and balanced modern standard arabic speech corpus. In: Second Workshop of Arabic Corpus Linguistics (WACL) (2013), https://api.semanticscholar.org/CorpusID:59726896
36. Sikasote, C., Anastasopoulos, A.: BembaSpeech: A speech recognition corpus for the Bemba language. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., Piperidis, S. (eds.) Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 7277–7283. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.lrec-1.790
37. Sikasote, C., Mukonde, E., Alam, M.M.I., Anastasopoulos, A.: BIG-C: a multimodal multi-purpose dataset for Bemba. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2062–2078. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.115, https://aclanthology.org/2023.acl-long.115
38. Sikasote, C., Siaminwe, K., Mwape, S., Zulu, B., Phiri, M., Phiri, M., Zulu, D., Nyirenda, M., Anastasopoulos, A.: Zambezi voice: A multilingual speech corpus for zambian languages. In: INTERSPEECH 2023. pp. 3984–3988 (2023). https://doi.org/10.21437/Interspeech.2023-1979
39. Tatman, R.: Gender and dialect bias in YouTube's automatic captions. In: Hovy, D., Spruit, S., Mitchell, M., Bender, E.M., Strube, M., Wallach, H. (eds.) Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. pp. 53–59. Association for Computational Linguistics, Valencia, Spain (Apr 2017). https://doi.org/10.18653/v1/W17-1606, https://aclanthology.org/W17-1606
40. Tatman, R., Kasten, C.: Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In: Proc. Interspeech 2017. pp. 934–938 (2017). https://doi.org/10.21437/Interspeech.2017-1746
41. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020), https://www.aclweb.org/anthology/2020.emnlp-demos.6
42. Zanon Boito, M., Besacier, L., Tomashenko, N., Estève, Y.: A study of gender impact in self-supervised models for speech-to-text systems. In: Interspeech 2022. pp. 1278–1282 (2022). https://doi.org/10.21437/Interspeech.2022-353
43. Zhang, Y., Zhang, Y., Patel, T., Scharenborg, O.: Comparing data augmentation and training techniques to reduce bias against non-native accents in hybrid speech recognition systems. In: 1st Workshop on Speech for Social Good (S4SG). pp. 15–19 (2022). https://doi.org/10.21437/S4SG.2022-4
44. Zhang, Y., Zhang, Y., Halpern, B., Patel, T., Scharenborg, O.: Mitigating bias against non-native accents. In: Interspeech 2022. pp. 3168–3172 (2022). https://doi.org/10.21437/Interspeech.2022-836