

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

Authors

Claytone Sikasote, Hussein Suleman and Jan Buys

University of Cape Town

South Africa



SACAIR2025



Introduction

- Training Automatic Speech Recognition (ASR) systems requires large amounts of training data to get good accuracy.
- Improved ASR due to advances in deep learning methods: **Multi-lingual pre-training**.
- State-of-the-art based on pre-trained models for low resource language (LRLs): **Fine-tuning**.

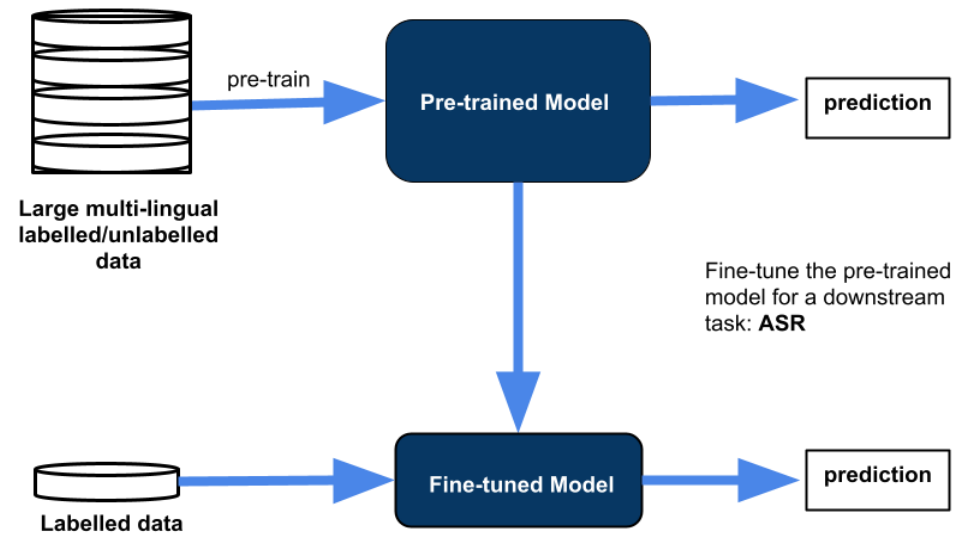
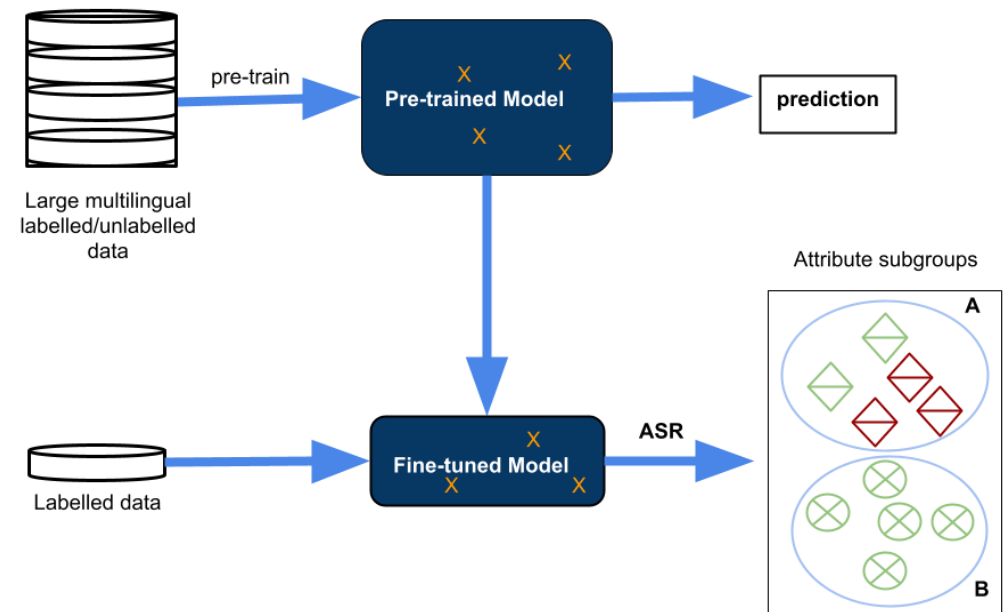


Fig. 1: Pre-training and fine-tuning process

Problem statement

- Fine-tuning increases the risk of introducing speaker attribute bias in the resulting ASR system.
- **Bias:** disparity in performance of an ASR system on speaker subgroups.
- **Speaker groups:** gender, age, non-native speakers, accents ++.



Biased/Unfair ASR system

Fig. 2: Biased/Unfair ASR system

Background

- Bias in artificial intelligence (AI) systems remains a challenge for developing inclusive technology system [1].
- Studies across a range of languages have investigated ASR systems for different forms of speaker attribute bias, such as gender, age, dialects, non-native & nationality [2].
- State of the art foundational speech models: XLSR, Massively Multilingual Speech (MMS), Whisper, and now the Omnilingual ASR [3].
- There are also African language-specialized models like AfriHuBERT [4].

Research Questions

- To what extent does fine-tuning pre-trained speech models impact gender bias in ASR systems fine-tuned on low resource African languages?
- Does the target language training data size affect gender bias in ASR systems fine-tuned on low resource African languages?

We investigate Gender Bias as a Binary Construct = {Male, Female} based on practical reasons.

Methodology: Pre-trained Speech Models

- **Massively Multilingual Model (MMS):** The Wav2Vec2-based pre-trained model
 - Trained approx. 500 hours of unlabelled audio data covering 1100 languages.
 - **Two variants:** 317 and 965 million parameter model.
 - We use the fine-tuned 1 billion parameter ASR MMS model.
 - It's integrated with language-specific adapters for efficient training and adaptation.
- **Whisper:**
 - Pre-trained on over 680K hours of weakly labeled data covering 98 languages.
 - **5 variants:** Base , Tiny, Small, Medium and Large whisper models .
 - We use the medium-sized model in our study.

Methodology: Target Languages and Datasets

Target Languages

- Bemba – spoken in Zambia/others
- Nyanja – spoken in Zambia/others
- Swahili – spoken in East Africa

Datasets Used

- **BembaSpeech (BS)** – Bemba
- **Bemba Image Grounded-Conversation (BIGC)** - Bemba
- **ZambeziVoice (ZV)** - Nyanja
- **CommonVoice (CV)** - Swahili

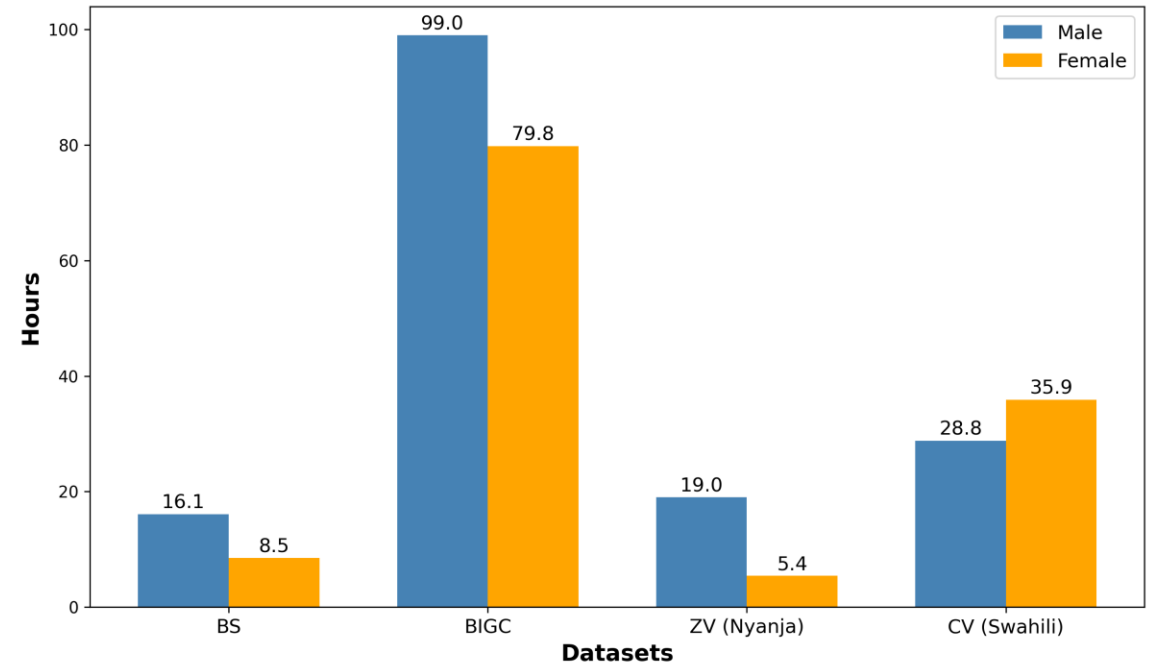
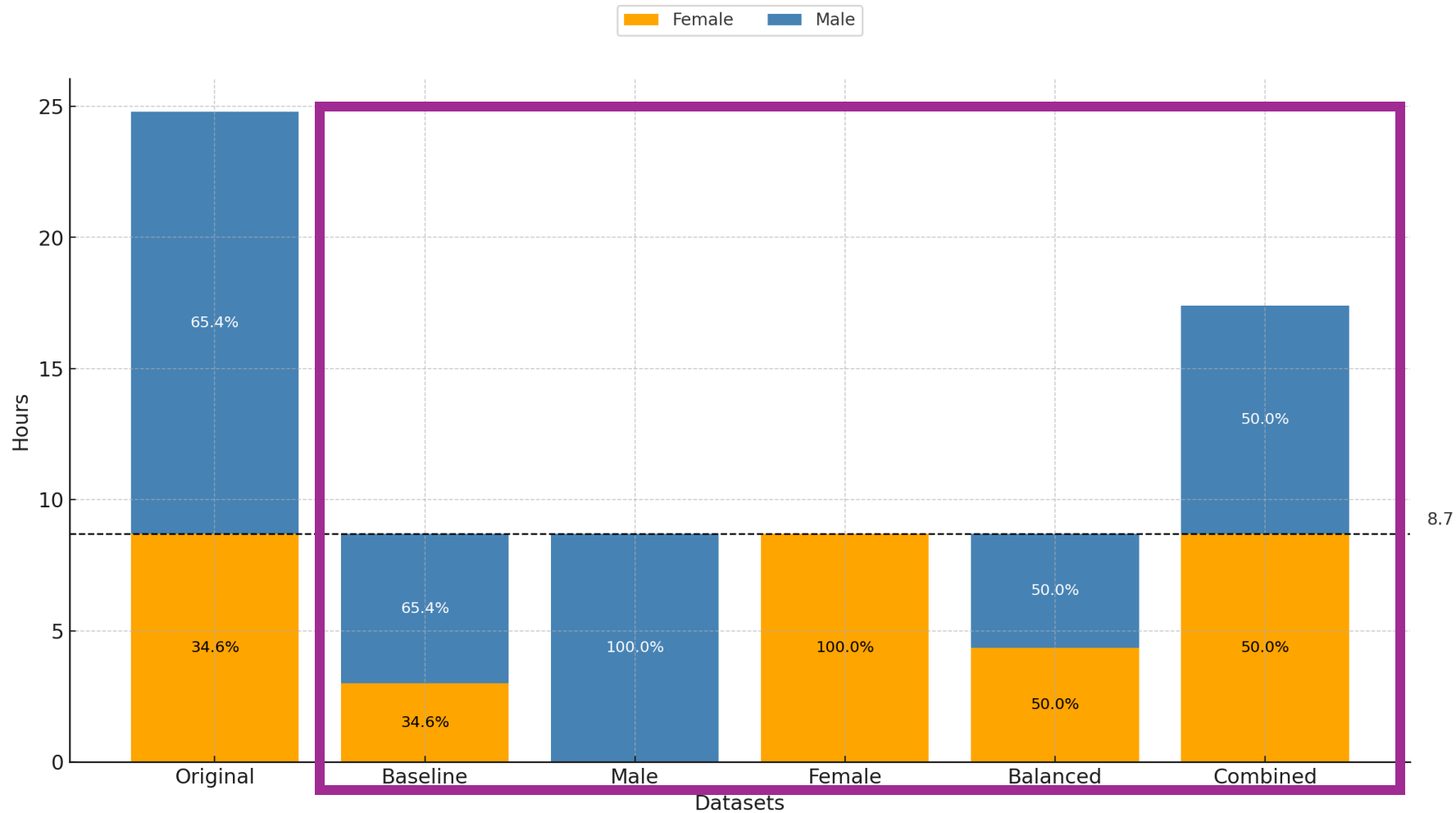


Fig. 3: Distribution of data by gender in original target datasets.

Methodology: Creating Training Datasets

Stacked Bar Chart: **BembaSpeech(BS) Datasets**



8.7 Fig. 4:
Balanced
datasets for
BembaSpeech

Methodology: Creating Training Datasets

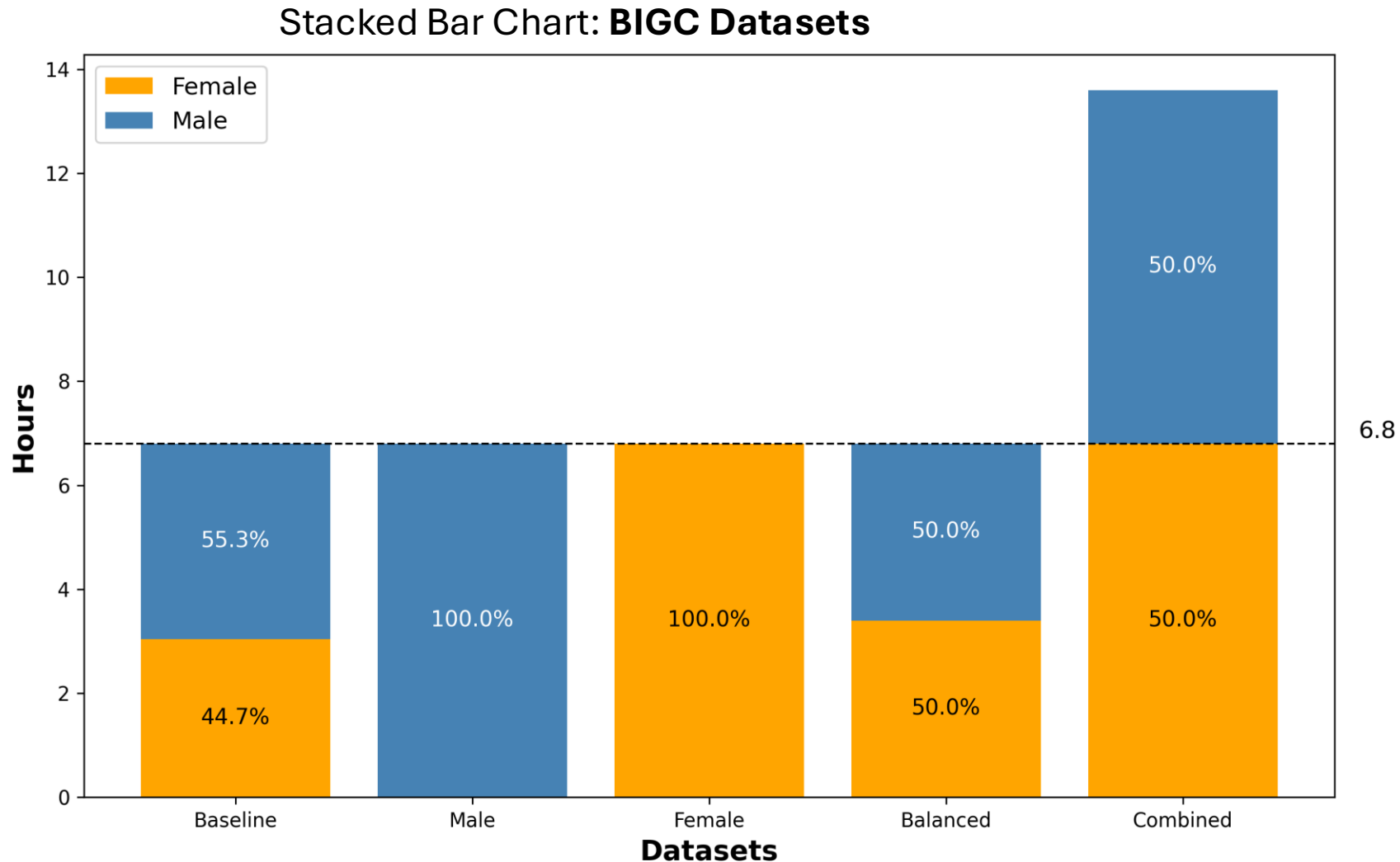


Fig. 5:
Balanced
datasets for
BIGC Dataset

Methodology: Creating Training Datasets

Stacked Bar Chart: Nyanja Datasets

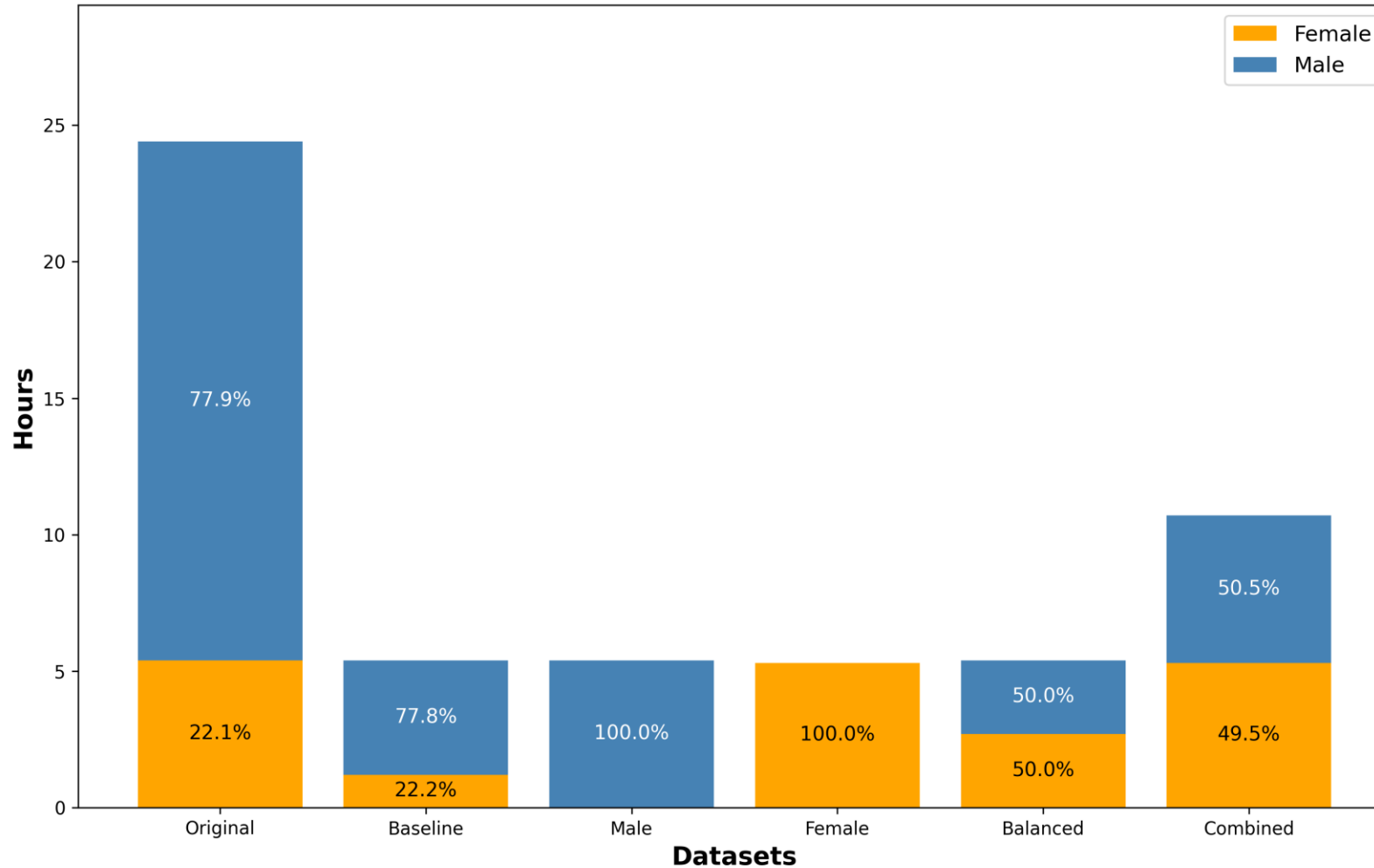


Fig. 6:
Balanced
datasets for
Nyanja Dataset

Methodology: Creating Training Datasets

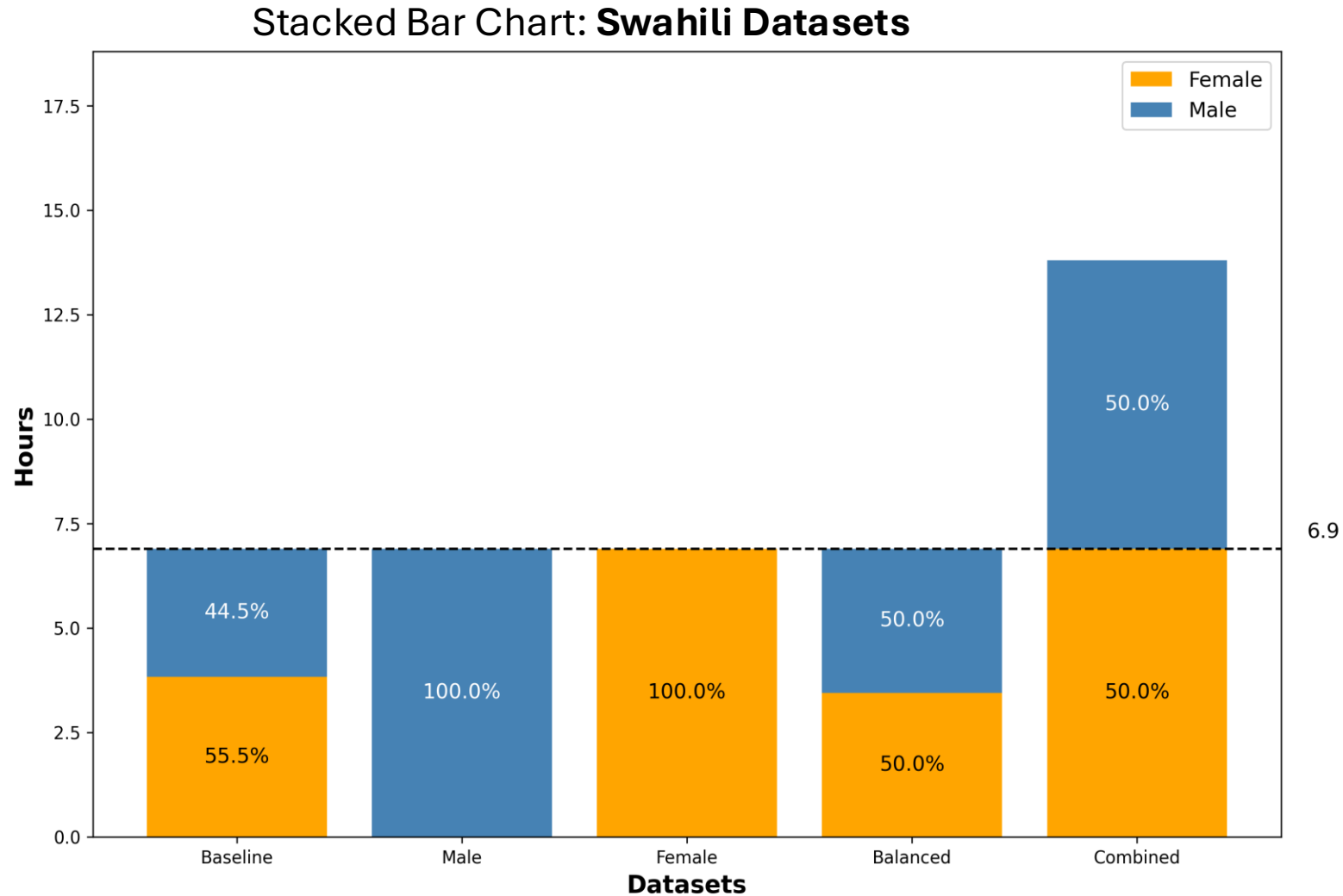


Fig. 7:
Balanced
datasets for
Swahili

Methodology: Experimental setup

- Evaluation criteria:
 - ASR: Word Error Rate (WER)
 - Gender Bias: difference of WERs between attribute subgroups.
- Fine-tuning strategy:
 - All models fine-tuned using the HF Transformer library with CTC objective.
 - Except for learning and batch sizes we inherited defaulting HF settings.
 - We trained models for 30 epoch with early stopping setting.
 - All models are trained on an A100 GPU.
- Evaluation sets:
 - Male-only, Female-only, Combined test sets.
 - No speaker overlap with training datasets

Results: Part 1 – ASR & Bias Analysis - MMS

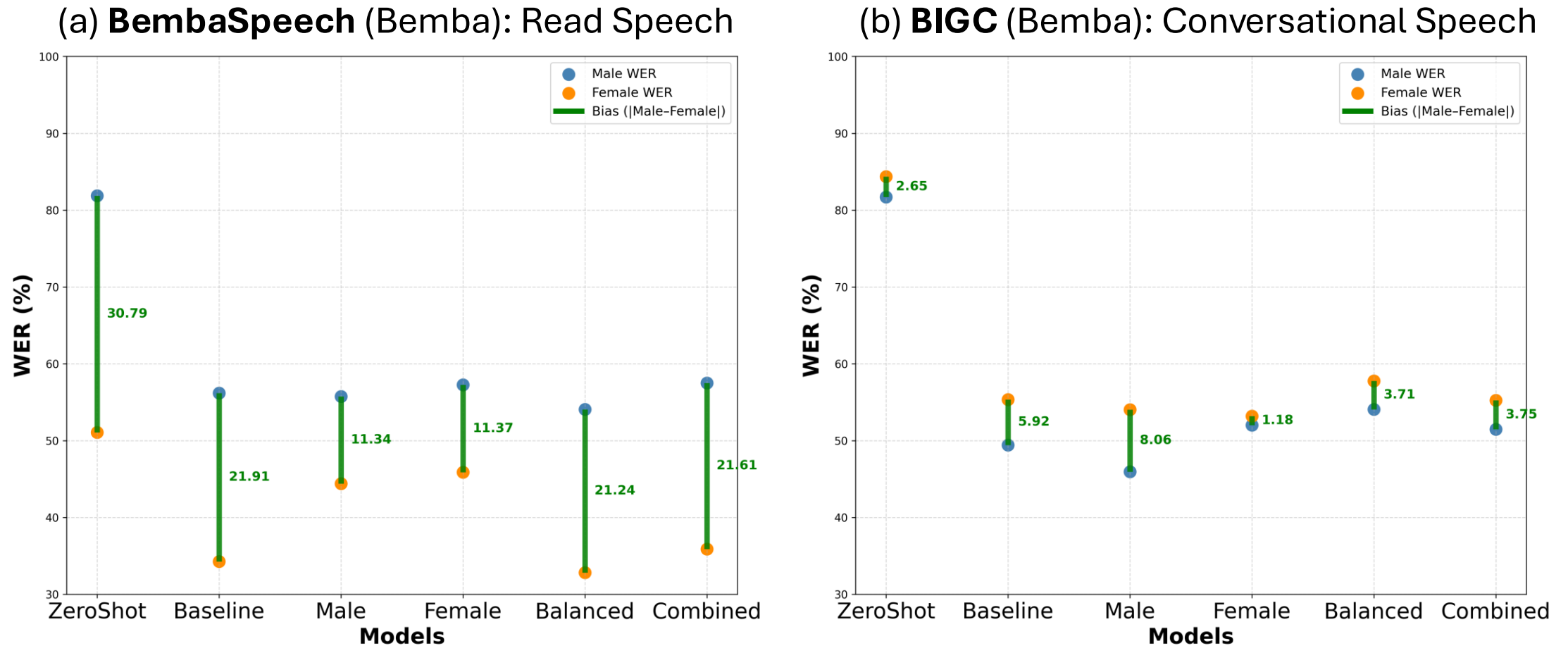


Fig. 8: Results of model fine-tuned BembaSpeech (BS) and BIGC datasets.

Results: Part 1 – ASR & Bias Analysis - MMS

Observations:

- All models recognize female speech more accurately than male speech.
- ZeroShot model exhibits higher bias than other models.
- Model fine-tuned on Baseline, Male, Female, Balanced and Combined outperform ZeroShot evaluation.
- Lower bias estimates compared to BS (Bemba)

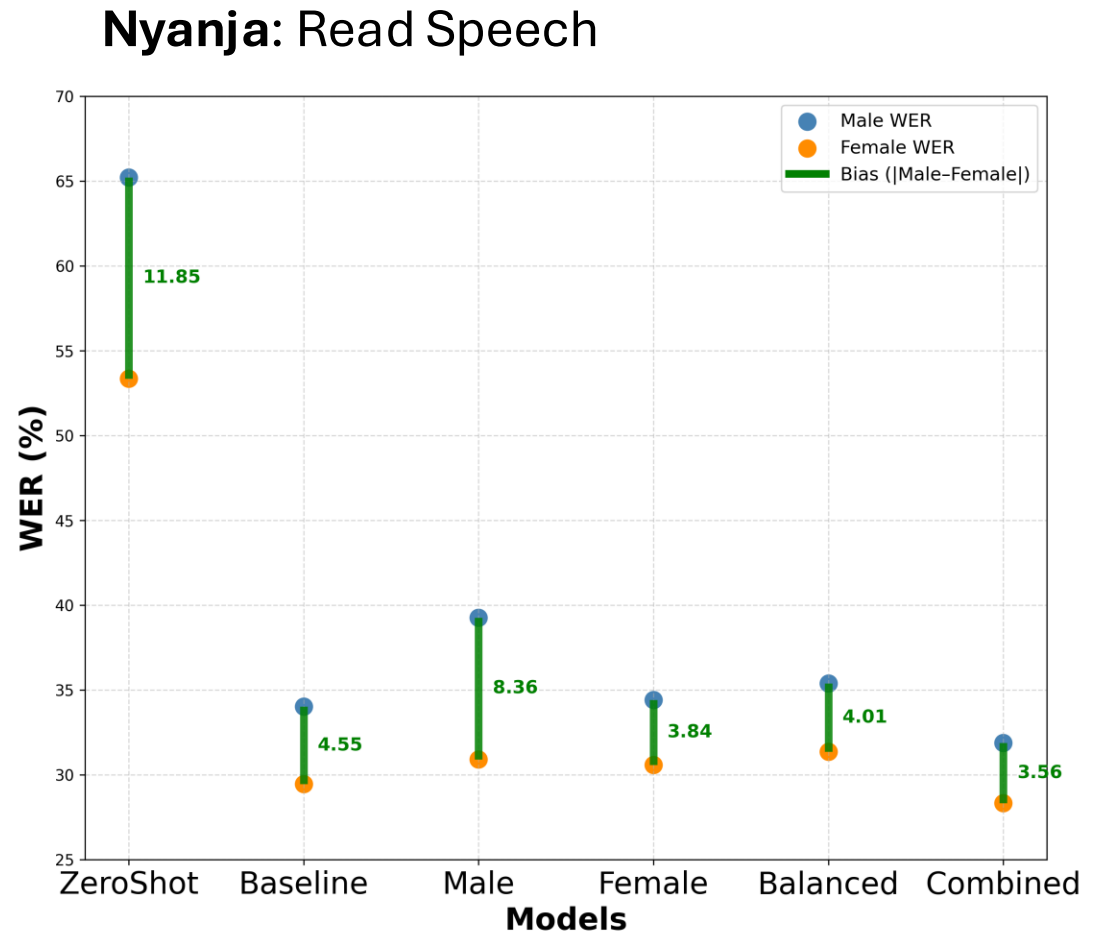


Fig. 9: Results for Nyanja

Results: Part 1 – ASR & Bias Analysis - MMS

Observation:

- All fine-tuned models recognize female speech more accurately than male speech.
- The ZeroShot-based model exhibit higher bias estimates.
- Relatively higher bias on gender-balanced-based models.

Swahili: Read Speech

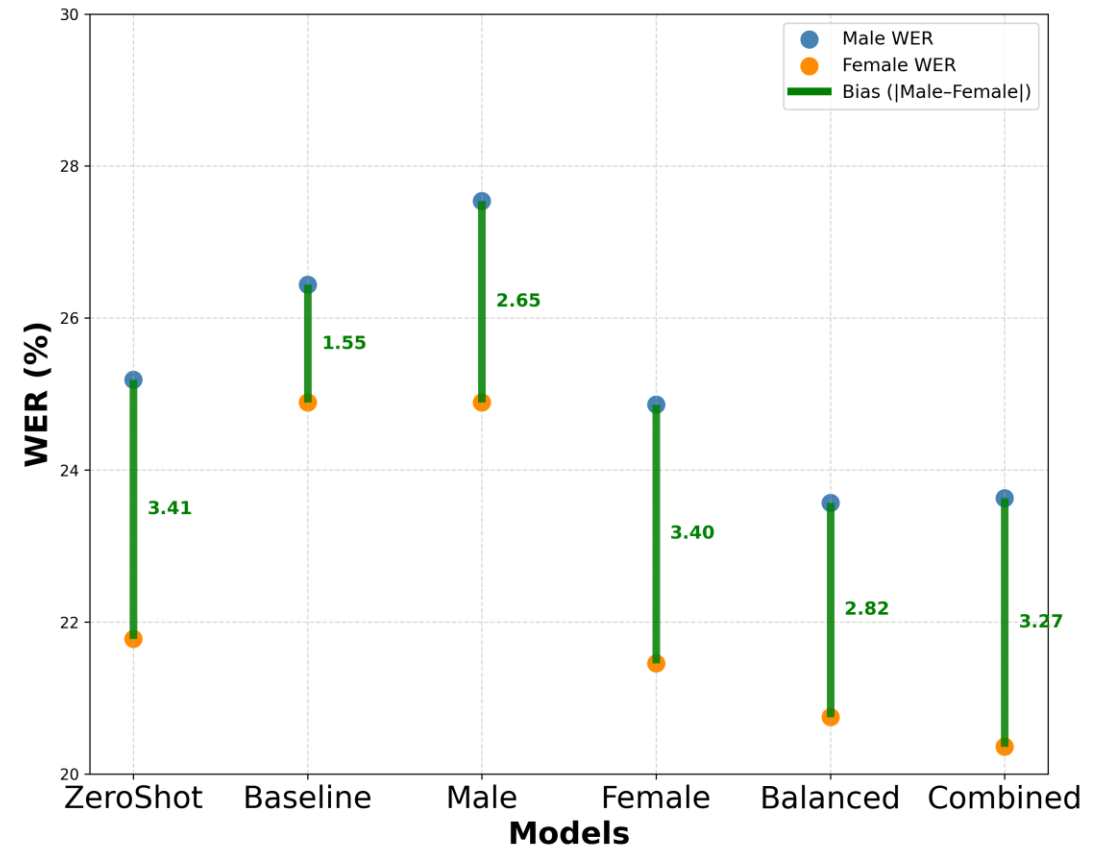


Fig. 10: Results for Swahili

Results: Part 2 - Gender Bias vs Training Data

Approach:

- Create different amount of gender-balanced training datasets.
- Fine-tune MMS and Whisper models on these datasets.
- Evaluate on **gender-specific test sets**.

Results

- There is no relationship between training data size and gender bias.

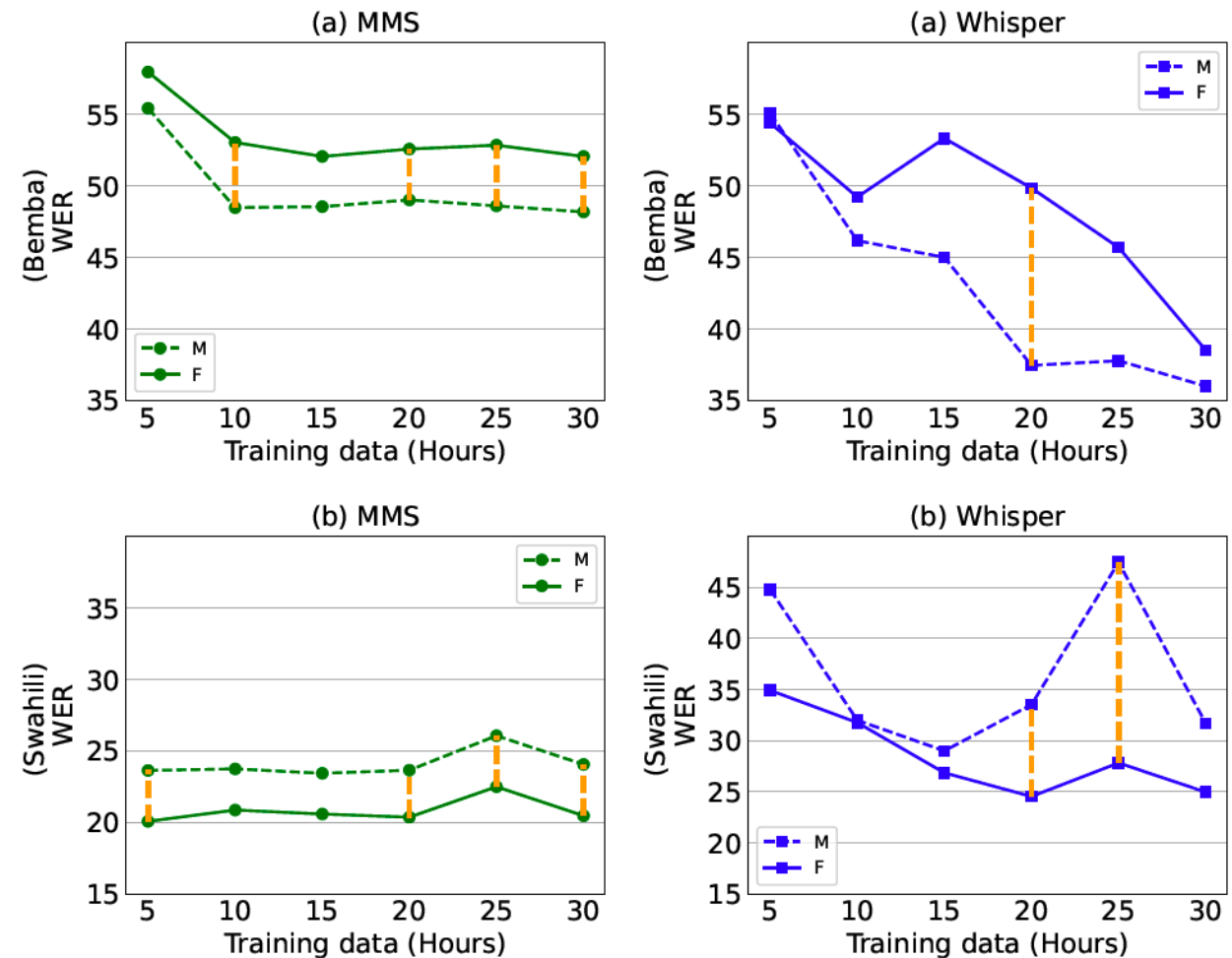


Fig. 11: Plot of training data size (Hours) vs model ASR performance (WER) on gender subgroup test sets.

Conclusion

- There is bias transfer from pre-trained speech models during fine-tuning
- The extent of bias in the fine-tuned model is language and dataset dependent.
- The speech-type of the target dataset has potential to determine which gender speech is recognized accurately.
- There is no relationship between training data size and gender bias.

Limitations & Future Work

- Limitations:
 - We covered only three languages: Bemba, Nyanja, and Swahili.
 - We evaluated only two dominant models: MMS and Whisper.
 - Limited attributes to investigate bias in target datasets.
 - We did not conduct error and linguistic analysis.
- Future Work:
 - Evaluate pre-trained models focused on African languages, such as AfriHuBERT.
 - Extend evaluation to other African languages
 - Investigate bias for other speaker attributes

References

- [1] 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>
- [2] Ngueajio, M.K., Washington, G.: 24th International Conference on Human-Computer Interaction, HCI 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
- [3] Omnilingual ASR Team et al.: Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages, 2025, <https://ai.meta.com/research/publications/omnilingual-asr-open-source-multilingual-speech-recognition-for-1600-languages>
- [4] 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>

Thank You!