The Balancing Act: Unmasking and Alleviating ASR Biases in Portuguese

Ajinkya Kulkarni, Anna Tokareva, Mohammed Rameez Qureshi, Miguel Couceiro

The Balancing Act: Unmasking and Alleviating ASR Biases in Portuguese

Ajinkya Kulkarni
MBZUAI, UAE
ajinkya.kulkarni@mbzuai.ac.ae

Anna Tokareva
University of Lorraine
anna.tokareva3@etu.univ-lorraine.fr

Rameez Qureshi
ADAPT Centre, Trinity College Dublin
rameez.qureshi@adaptcentre.ie

Miguel Couceiro
University of Lorraine, CNRS, LORIA
miguel.couceiro@loria.fr

Abstract

In the field of spoken language understanding, systems like Whisper and Multilingual Massive Speech (MMS) have shown state-of-the-art performances. This study is dedicated to a comprehensive exploration of the Whisper and MMS systems, with a focus on assessing biases in automatic speech recognition (ASR) inherent to casual conversation speech specific to the Portuguese language. Our investigation encompasses various categories, including gender, age, skin tone color, and geo-location. Alongside traditional ASR evaluation metrics such as Word Error Rate (WER), we have incorporated p-value statistical significance for gender bias analysis. Furthermore, we extensively examine the impact of data distribution and empirically show that oversampling techniques alleviate such stereotypical biases. This research represents a pioneering effort in quantifying biases in the Portuguese language context through the application of MMS and Whisper, contributing to a better understanding of ASR systems' performance in multilingual settings.

1 Introduction

Conversational Artificial Intelligence (AI) has become increasingly integrated into everyday applications over the past few years. The history of previous broad technologies shows that despite temporary challenges, restructuring the economy around innovative technologies offers significant long-term benefits (Mühleisen, 2018). This asks for fair AI solutions that can connect people from different backgrounds, and that enables universal access to technology. In the context of human-machine interactions through spoken language, Automatic Speech Recognition (ASR) facilitates smooth information exchange within various conversational AI applications, including machine translation, sentiment analysis, and question-answering systems (Bangalore et al., 2005).

The significance of spoken language in our daily lives emphasizes the need for ASR systems to accommodate the various forms of human communication. It is thus vital that ASR systems can adeptly manage this diversity, as it is crucial for enabling smooth and inclusive communication across a wide range of situations and people, and extending the use of ASRs in domains such as emergency services, home automation, and navigation systems. To accommodate fairness and transparency requirements it is paramount to examine the prevailing biases within various subgroups towards fair ASR systems.

Over the past few years, there has been a growing research community examining biases in automatic speech recognition (ASR) systems (Koenecke et al., 2020; Tatman, 2017; Tatman and Kasten, 2017; Harwell, 2018; Lima et al., 2019; Blodgett et al., 2020). This research has primarily focused on assessing the impact of disparities related to gender, age, accent, dialect, and racial meta-attributes. (It is worth mentioning that most of these features are considered sensitive according to legal protection against discrimination, *e.g.,* in the U.S.[1] and in Europe[2].) However, the majority of these studies have been carried out on monolingual ASR systems for the English language, with only a limited number of studies addressing bias detection in non-English languages.

In the study conducted in (Feng et al., 2021, 2024), researchers examined the (Hidden Markov Model) HMM- Deep Neural Network (DNN) ASR system to assess biases related to gender, age, and accents in the context of the Dutch language. They then proposed the use of data augmentation and vocal tract length normalization techniques to alleviate these biases in Dutch ASR systems (Pa-

[1] https://www.whitehouse.gov/ostp/ai-bill-of-rights/#applying

[2] https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

tel and Scharenborg, 2023). Another study centered on French broadcasting speech, aimed to uncover gender biases and revealed that the underrepresentation of specific gender categories could result in bias in HMM-DNN ASR performance, regardless of gender identity (male, female, or other) (Adda-Decker and Lamel, 2005; Garnerin et al., 2019). Furthermore, it emphasized the importance of a systematic examination of demographic imbalances present in datasets.

For Arabic ASR system, which were developed using Carnegie Mellon University Sphinx 3 tools[3], an investigation was conducted to understand the impact of gender, age, and regional factors on performance (Sawalha and Shariah, 2013). While these studies laid the foundation for quantifying biases, there remains a scarcity of research on ASR systems trained with large amounts of multilingual data, even though they consistently achieve state-of-the-art performance levels.

The emergence of computational resources enabled the acceleration of the development of large pre-trained acoustic models, resulting in unified frameworks with multilingual capabilities. These frameworks are often built upon transformer networks and prominently use the Wav2vec 2.0 (Baevski et al., 2020) framework. As a consequence, there has been a significant push to create multilingual ASR systems(Li et al., 2022; Alec Radford, 2023; Zhang et al., 2023; Pratap et al., 2023), extending their applicability to more than 100 languages, including those with limited linguistic resources. Meta AI's MMS system (Pratap et al., 2023) conducted an evaluation that included the assessment of gender and language biases using the FLEURS dataset (Conneau et al., 2022). However, there is still a need for a deeper understanding of the comparative differences among various multilingual ASR systems when it comes to quantifying potential biases.

To explore the biases present in multilingual ASR systems trained on extensive speech data, we investigated variants of OpenAI's Whisper ASR system (Alec Radford, 2023) and Meta AI's MMS ASR system (Pratap et al., 2023), both of which have achieved state-of-the-art performance levels. In addition, we selected the Casual Conversation Dataset version 2 (CCD V2) to quantify biases and assess the fairness of these system performances in the context of the Portuguese language (Porgali et al., 2023). Our study takes into account a diverse spectrum of categories, including age groups, gender, geographical locations, and skin tones. The consistency in textual content across all CCD V2 recordings establishes a robust basis for the efficient evaluation of system performance across a broad array of categories. Only a limited number of studies have delved into the influence of state-of-the-art multilingual ASR systems on domain-specific ASR tasks. For example, these studies have explored code-switching between languages using systems like Whisper and MMS (Kulkarni et al., 2023), or they have examined the effects of ASR errors on discourse models among groups of students in noisy, real-world classroom settings between Whisper and Google ASR system (Cao et al., 2023).

More often, an imbalanced distribution of evaluation data across various sub-categories can result in an inadequate analysis of the evaluation process itself. Therefore, we explore two resampling methods, namely, *naïve* and *Synthetic Minority Oversampling Technique* (SMOTE)(Chawla et al., 2002), to ensure a balanced data distribution across each subgroup when quantifying the biases. In the assessment of ASR systems, our primary choice of metrics includes Word Error Rate (WER) and Character Error Rate (CER)[4]. Interestingly, we observe that oversampling techniques can alleviate performance disparities between certain subgroups.

The structure of the paper is as follows. In Section 2, we provide an overview of the Casual Conversation dataset, which is utilized to quantify biases in multilingual ASR systems in the Portuguese language. We described the specifics of the MMS ASR system and the variants of Whisper ASR systems along with the evaluation protocol in Section 3. We outline results along with an analysis on various categories to to quantify biases in Section 4, along with the corresponding evaluation methodologies. Section 5 details the discussion, and we draw our conclusions in Section 6 along with potential directions for future work.

The **main contributions** of this paper are as follows:

1. It presents the first study on analyzing disparities within multilingual ASR systems focused

[3] https://www.cs.cmu.edu/~archan/sphinxInfo.html

[4] In this paper, we only include the WER results. The CER results are provided in https://biasinai.github.io/asrbias/.

on the Portuguese language.

2. It emphasizes the critical significance of data distribution among sub-categories by employing oversampling techniques.

3. It illustrates the comparative distinctions between Whisper ASR and MMS ASR, and examines the impact of model parameters on the development of an efficient system design.

4. In addition to gender and age groups, it investigates skin tone and geo-location as criteria to measure inter-racial biases.

2 Dataset Description

The CCD V2 dataset is open-source and can be accessed through the Meta AI website[5]. It represents the speech of 5,567 unique speakers from various regions, including India, the United States of America, Indonesia, Vietnam, Brazil, Mexico, and the Philippines. This compilation results in five audio samples per individual, yielding a total of 26,467 video recordings. The dataset encompasses seven self-labeled attributes, including details about the speaker's age, gender, native and secondary languages or dialects, disabilities, physical characteristics, and adornments, as well as geographic location. Additionally, it features four other characteristics: two skin tone scales (Monk Skin Tone (Monk, 2019) and Fitzpatrick Skin Type (Molina et al., 2020; Ash et al., 2015)), voice timbre, the speaker's activity, categorized as gesture, action, or appearance, and details about the recording setup, which covers video quality, background environment, and video configuration. For Monk skin tone scale-10 only one sample was available for Portuguese language. Therefore, in order to avoid skewed comparison between skin-tone scales using Monk skin tone, we only conducted a study using Fitzpatrick skin type.

The CCD V2 comprises 354 hours of recordings where speakers responded to specific questions in a non-scripted manner and 319 hours of recordings in which individuals read passages from F. Dostoyevsky's "The Idiot", translated into various languages. Throughout this paper, we utilized scripted recordings for the Portuguese language. As each scripted recording had the same textual content and phonetic variations, it enables the examination of meta-attributes leading to performance differences. For more comprehensive details of CCD V2 and the dataset design process, please refer to the works published in (Porgali et al., 2023) and (Hazirbas et al., 2021).

In the context of assessing the fairness of ASR systems, we focused primarily on a subset of scripted recordings, with a strong emphasis on the Portuguese language. In this study, we concentrated on four annotated labels: gender, age, Fitzpatrick scale, and geographic location. To simplify our analysis, we categorized speakers into seven age groups: 18-24, 25-30, 31-36, 37-42, 43-50, 51-60, and 61+. After the initial analysis of the evaluation sample distribution for each sub-category, we observed imbalanced distributions among various subgroups. We thus explored resampling strategies to ensure that biases are not introduced into the computed results due to imbalanced distributions across subgroups.

3 Empirical study

In this empirical study, we initiate our investigation by conducting a thorough analysis of the influence of various sampling techniques on performance disparities within multilingual ASR systems for Portuguese. Additionally, in Section 3.1, we first present the ASR systems employed in this research. Subsequently, we outline the evaluation protocol and data preparation in Section 3.2.

3.1 ASR Systems

This study centers around the utilization of state-of-the-art, open-source multilingual ASR systems, specifically Whisper and the Multilingual Massive Speech Systems. Both of these systems have demonstrated their efficacy in a range of speech-processing tasks, including audio classification, speech translation, and text-to-speech synthesis. They have been trained on extensively large-scale multilingual datasets using self-supervised and multi-task learning techniques, enabling support for over 100 languages.

3.1.1 Whisper

Whisper (Alec Radford, 2023) is a robust speech recognition model presented by OpenAI[6] in 2022. Whisper is trained using a multitask learning on 680,000 hours of labeled multilingual recordings

[5]https://ai.meta.com/datasets/casual-conversations-v2-dataset/

[6]https://openai.com/research/whisper

collected from the Internet, along with the corresponding transcriptions filtered from machine-generated ones. In total 96 languages are covered by approximately 117,000 hours of audio data, making Whisper a powerful tool for multilingual speech recognition.

Whisper incorporates the Transformer encoder-decoder architecture (Vaswani et al., 2017) with the implementation of multitask learning techniques allowing language identification, multilingual speech transcription, along with word-level timestamps. The input audio is split into thirty-second chunks, which makes the transcription of long recordings more effective. In the Whisper framework, the encoder processes log Mel spectrogram inputs, generating relevant features for the decoder. The decoder, in turn, consumes these encoder features, positional embeddings, and a sequence of prompt tokens. Subsequently, it produces the transcribed text corresponding to the input speech.

Whisper has different variants based on model parameter sizes such as Tiny (39 Million), Base (74 Million), Small (244 Million), Medium (769 Million), Large (1550 Million), and Large-v2 (1550 Million). Whisper models are primarily divided into two categories based on languages and tasks: English-only models and multilingual models. In this paper, we incorporated Medium, Large, and Large-v2 variants of Whisper.

3.1.2 Massively Multilingual Speech system

In 2023, Meta AI released the Massively Multilingual Speech (MMS) project, as documented in (Pratap et al., 2023), expanding its language support to encompass over 1000 languages for various speech processing applications. The primary components of the MMS system include a novel dataset derived from publicly accessible religious texts and the adept use of cross-lingual self-supervised learning. The MMS project encompasses various tasks, such as speech recognition, language identification, and speech synthesis. MMS is built upon the Wav2Vec 2.0 (Baevski et al., 2020) architecture and has undergone training through a combination of cross-lingual self-supervised learning and supervised pre-training for ASR. It incorporates language adapters that can be dynamically loaded and interchange during inference, featuring multiple Transformer blocks, each augmented with a language-specific adapter.

The authors compiled two datasets using texts from the New Testament and the Bible, along with recordings of readings of these religious texts available on the Internet. The labeled dataset (MMS-lab) comprises 1,306 audio recordings of New Testament readings in 1,130 languages, resulting in 49,000 hours of data and approximately 32 hours of data per language. The audio underwent several alignment stages, including training several alignment models and a final filtering of noisy or paraphrased data. The unlabeled dataset (MMS-unlab) contains 9,345 hours of audio and includes recordings collected from the Global Recordings Network, organized into 3,860 languages. The MMS system is available in two variants based on model parameters, with 317 million and 965 million parameters. For this study, we utilized the MMS system with 965 Million model parameters.

3.2 Preprocessing and evaluation processes

In this subsection, we will first outline the preprocessing steps employed to prepare the evaluation dataset using CCD V2 for Portuguese. We will explain the sampling methods for analyzing biases within sub-categories and subsequently discuss the evaluation measures used to assess disparities among these sub-categories.

3.2.1 Handling imbalance

Imbalanced evaluation data can have a detrimental effect on the results, making it challenging to discern meaningful distinctions between the groups being compared. From Table 1, we observe that initially collected samples for Portuguese have unbalanced distributions across several categories, which may impact the assessment of ASR systems towards measuring disparities towards underrepresented classes. Therefore, we opted for data balancing approaches, specifically focused on oversampling, and subsequently compared the results.

It is also worth mentioning that after preliminary analysis of ASR systems results, we observed that the Portuguese subset of the CCD V2 dataset contains audio recordings named "Portuguese scripted" but representing the speech of people speaking on various topics but not reading the passage from Dostoevsky's novel. This might have been a mistake during the compilation of the CCD V2 dataset. These samples were deleted from our evaluation data since the WER for the corresponding transcriptions was exceptionally high and negatively affected the overall performance.

At first, we used Naïve sampling (Naïve) based on the 'gender' category since the WER values

	Gender		Fitzpatrick scale						Age Groups							Geo-location										
	Male	Female	T.1	T.2	T.3	T.4	T.5	T.6	18-24	25-30	31-36	37-42	43-50	51-60	61+	MA	MT	RN	GO	PI	RS	RJ	SP	PE	PR	MG
Initial	240	500	11	192	289	159	72	17	83	201	164	137	103	44	8	9	27	25	11	7	28	130	379	38	55	31
Naïve	1019	1009	25	681	743	345	164	70	293	282	297	293	283	274	285	18	47	39	24	34	70	409	1110	59	119	99
SMOTE	4443	4443	1925	1893	2630	1132	322	984	1014	2918	1703	1236	978	458	579	892	854	845	841	830	805	796	787	769	744	723

Table 1: Statistical representation of samples for demographic categories across Initial, Naïve, and SMOTE datasets. The abbreviations for 'Geo-location' are as follows: RN - Rio Grande do Norte, SP - Sao Paulo, RS - Rio Grande do Sul, GO -Goias, MT - Mato Grosso, PR - Parana, RJ - Rio de Janeiro, MG - Minas Gerais, PI - Piaui, PE - Pernambuco, MA - Maranhao. The abbreviations for 'Fitzpatrick scale' are as follows: T.1 - type i, T.1 - type ii, T.1 - type iii, T.1 - type iv, T.1 - type v, T.1 - type vi.

for this category appeared to differ significantly. We achieved data balance by randomly duplicating instances until we had an approximately equal number of male and female records. However, we found that naïve sampling did not improve the balance of the other categories. Therefore, we turned to the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) in the final stage.

The SMOTE algorithm aims to tackle the issue of imbalanced data by creating synthetic observations for minority classes. It does not simply repeat the existing samples but rather creates similar examples that improve performance accuracy. It starts with choosing an instance in the minority class and computing the difference of feature vectors with neighboring observations. After that, the algorithm defines a region of k nearest neighbors around the selected instance. Next, the algorithm calculates the difference between observations and multiplies the difference vector by a random number from the range (0, 1), thus having a new synthesized sample. We do the resampling for every category one by one assuming the improvement in results.

The statistics for evaluation data compiled using oversampling techniques along with initial samples are shown in Table 1. The average duration of each sample used for the evaluation of multilingual ASR systems corresponds to 2 minutes with the same textual content. Therefore, the robustness of ASR systems to long-form audio is an important consideration in the development and deployment of ASR technology.

3.2.2 Evaluation strategy

For the evaluation of both models, we use the Word Error Rate (WER), a standard metric for ASR. The Word Error Rate depicts the percentage of incorrectly recognized words and is calculated as follows:

$$WER = \frac{S + D + I}{N}, \quad (1)$$

where S stands for number of substitutions, D for the number of deletions, I is the number of insertions, and N for the number of words in the reference transcription. In the current paper, we report the WER for comparison purposes with the literature, and we also report the Character Error Rate (CER) in https://biasinai.github.io/asrbias/. This allows us to compare results objectively and to identify performance biases in the 4 ASR systems.

Method	W-L	W-L-V2	W-M	MMS
Initial	0.00022	0.00018	0.0011	0.195
Naïve	2.07e-17	1.54e-17	1.45e-11	0.177
SMOTE	0.676	0.603	0.778	0.563

Table 2: p-values for Whisper ASR variants and MMS for the Gender category across Initial, Naïve and SMOTE datasets. Whisper ASR variants are indicated as, Whisper-Large (W-L), Whisper-Large-V2 (W-L-V2), and Whisper-Medium (W-M).

4 Results and Analysis

In this section, we present a comprehensive analysis of Word Error Rate (WER) within distinct categories as provided by CCD V2. These categories include gender (Section 4.1), skin tone (Section 4.2), age groups (Section 4.3), and geo-location (Section 4.4). As previously mentioned, our experimentation involved the use of three Whisper ASR variants: Medium (769 million parameters), Large (1550 million parameters), and Large-v2 (1550 million parameters)[7] (which maintains the same parameter count but benefits from extended training with regularization). Additionally, we utilized the MMS ASR system[8] with 965 million parameters.

4.1 Gender analysis

We illustrate the performance of ASR systems for the Portuguese language, on the gender subgroups 'Male' and 'Female'. From Figure 2, we observe a subtle gender bias when examining the Whisper ASR variants, which favors males in both the Initial and naïve sampling techniques. However, the use of SMOTE sampling results in a more balanced ASR performance between the gender sub-

[7] https://huggingface.co/openai/whisper-large-v2

[8] https://huggingface.co/facebook/mms-1b-all

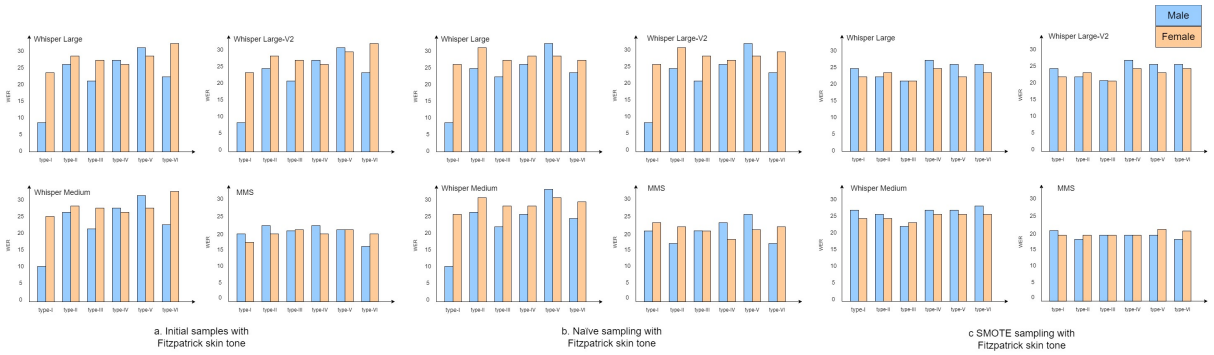

Figure 1: Bar plots depicting Whisper and ASR performance across the Fitzpatrick skin-tone scale, ranging from type-I to type-VI, for both male and female genders, with results for initial samples, naïve sampling, and SMOTE sampling

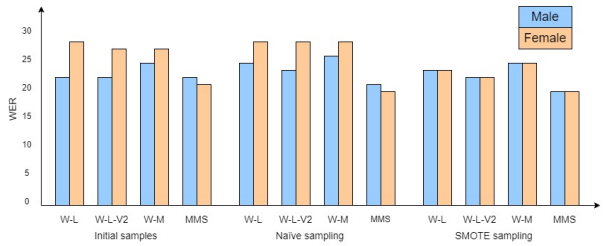

Figure 2: Bar-plots demonstrating performance of multilingual ASR systems using Whisper ASR variants and MMS for impact on male and female genders using WER under three sampling methods, initial, naïve and SMOTE. Whisper ASR variants are indicated as, Whisper-Large (W-L), Whisper-Large-V2 (W-L-V2), and Whisper-Medium (W-M).

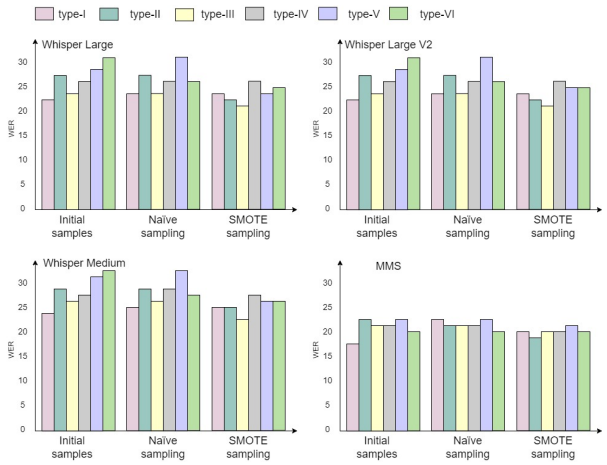

Figure 3: Bar-plots illustrating the distribution of mean WER for Fitzpatrick skin tone scales across Initial, naïve, and SMOTE sampling methods.

groups. Notably, the MMS system outperforms the Whisper ASR variants, exhibiting comparatively balanced WER across both genders. As illustrated in Figure 2, we observe the absence of significant performance disparities between male and female genders.

In addition to analyzing WERs, we also conducted a p-value analysis to assess the statistical significance of gender-related differences. In the examination of Table 2, we observed that the p-values for Whisper ASR variants applied to initial samples and Naïve sampling fell below the significance threshold of 0.05. This suggests that statistically significant differences exist between male and female gender categories in these cases. Conversely, the p-value statistics for the MMS approach consistently exceeded 0.05, indicating that there are no significant performance variations across both genders regardless of the sampling method. Regarding SMOTE sampling, the p-values for all ASR systems exceeded the 0.05 threshold, signifying evidence of mitigating gender biases in this context.

After this, we extended our study of ASR systems with the distribution of WER performances concerning skin tone as measured by the Fitzpatrick skin type and gender. This examination is depicted in Figure 1. Significant disparities are evident across different skin tone types between male and female individuals. Specifically, within the Whisper ASR variants, notable performance differences are observed for skin-tone type-I and type-VI. In these cases, the male subgroup exhibits better WER compared to the female subgroup, particularly in the context of initial samples and naïve sampling approaches. Moreover, the MMS ASR system demonstrates a relatively even distribution of WER across all skin-tone types and outperforms all variants of the Whisper ASR. It is worth highlighting that, across all the ASR systems under examination, the use of SMOTE sampling has consistently played a role in mitigating performance disparities, leading to more balanced outcomes across gender

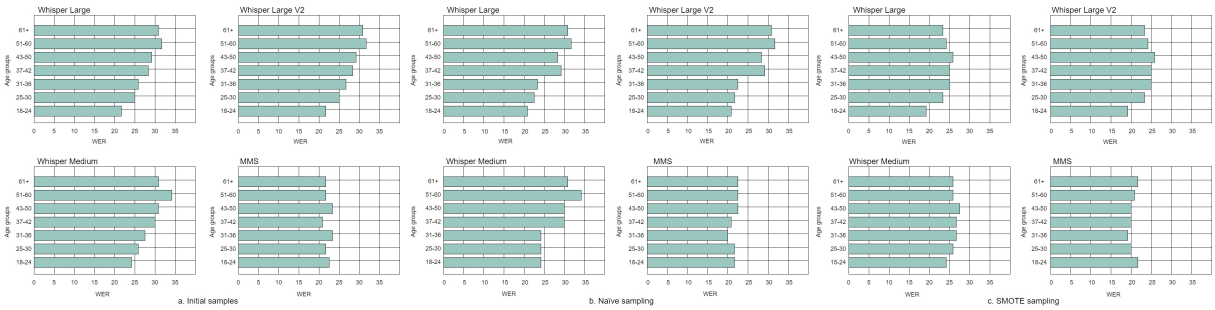

Figure 4: Bar-plots illustrating distribution of WER for age groups categorized into five sub-sets (18-24, 25-30, 31-36, 37-42, 42-50, 51-60, 61+) across initial, naïve and SMOTE sampling methods.

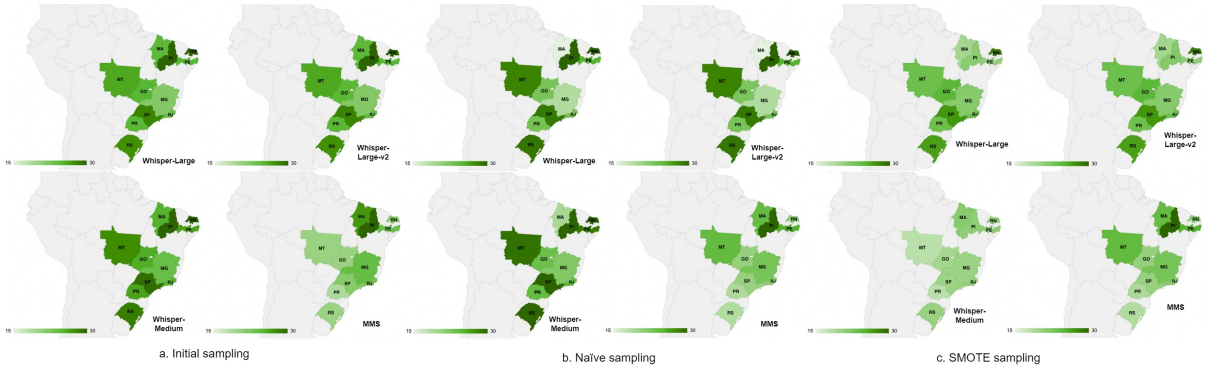

Figure 5: The visualization of mean WER distribution in each Portuguese state. The abbreviations of states are as follows: RN - Rio Grande do Norte, SP - Sao Paulo, RS - Rio Grande do Sul, GO -Goias, MT - Mato Grosso, PR - Parana, RJ - Rio de Janeiro, MG - Minas Gerais, PI - Piaui, PE - Pernambuco, MA - Maranhao.

subgroups.

4.2 Skin-tone analysis

We also examine the impact of ASR performance within sub-categories using categorized by Fitzpatrick skin tone type, without conditioning on other meta-attributes. Figure 3 shows the relative performance variations across various sampling techniques applied to ASR systems. Notably, we observe that individuals with skin types I to III demonstrate comparatively better WER than those with skin type IV. This observation sheds light on potential racial biases in ASR systems, where greater skin-type variations often indicate darker skin colors.

However, amidst these disparities, the MMS ASR system stands out with evenly distributed WER measures across all skin-type scales. When assessing the differences introduced by sampling approaches, initial samples, and naïve sampling reveal disparities among skin-tone subgroups. In contrast, the consistent use of SMOTE sampling proves effective in mitigating discrepancies across all the ASR systems under investigation.

4.3 Age group analysis

In Figure 4, we present an age group analysis of the Portuguese language for ASR systems using three different sampling techniques: initial samples, naïve sampling, and SMOTE sampling. Across all the sampling methods, the MMS ASR system consistently maintains WER measures below 25% for all age groups, exhibiting a relatively even distribution of WER values. In contrast, the Whisper ASR variants demonstrate disproportionate WER measures, particularly noticeable between the age groups of 18-36 and 36+. Moreover, the performance of the Whisper ASR degrades as age groups increase. However, the utilization of SMOTE sampling significantly improves the WER of the Whisper systems, bringing it to an overall 25%.

This distinctively highlights the positive impact of SMOTE sampling in reducing performance disparities across various age groups for both the Whisper and MMS ASR systems.

4.4 Geo-location analysis

Figure 5 provides a comprehensive examination of the impact of different sampling techniques on ASR performance disparities across various regions

in Brazil. Notably, when considering the Whisper ASR system, regions such as São Paulo (SP), Piauí (PI), Rio Grande do Norte (RN), and Rio Grande do Sul (RS) are notably affected by performance differences, regardless of whether initial samples or naïve sampling methods are employed. These regions exhibit significant variations in WER compared to other regions. Overall, the MMS ASR system displays a more even distribution of evaluation measures across all sampling approaches and generally outperforms the Whisper ASR variants. Furthermore, it is notable to highlight that, despite observing proportionate WERs across most regions in Brazil, the MMS ASR system experiences a decline in performance specifically in the Piauí (PI) region for all sampling approaches.

Even after the application of SMOTE sampling, the Whisper ASR variants continue to exhibit consistently higher WER values in the Rio Grande do Norte (RN) region. However, SMOTE sampling effectively mitigates WER discrepancies in the Piauí (PI) region. This underscores the distinct challenges posed by regional variations in ASR performance and underscores the potential of SMOTE sampling in addressing these disparities.

5 Discussion and limitations

Our results reveal that all 4 models show mild WER performance disparities when considering the individual subgroups of the categories 'Gender', 'Age', 'Skin Tone Color', and 'Geo-location', with a consistently better performance of the MMS model over the three Whisper models. However, when analyzing the gendered subcategories of 'Age', 'Skin Tone Color', and 'Geo-location', we observe significant differences in WER, with a noticeable bias that privileges the 'Male' subgroup; see additional results in `https://biasinai.github.io/asrbias/`.

Our study also shows that oversampling approaches can alleviate these disparities between the two gender subgroups. This is particularly evident in Figure 2, where WER performances are balanced for the 'Male' and 'Female' subgroups over the 4 models considered. The same trend was also observed for the other gendered categories and with respect to the Character Error Rate (CER) in the link provided earlier. The study shows that performances of Whisper variants demonstrate higher sensitivity to the number of model parameters, whereas the MMS system, despite having 40% fewer parameters than Whisper Large, exhibits better robustness over the various categories.

Despite promising, these results naturally ask for similar comparisons with respect to other performance and bias metrics. Another limitation of our study is that it was carried out solely on the CCD V2. In (Meyer et al., 2020), the Artie Bias Corpus is curated as a subset of the Mozilla Common Voice corpus. It includes demographic tags for age, gender, and accent, which allows for the examination of disparities in the English language. It is imperative to construct bias-focused datasets using publically available resources for Portuguese.

Furthermore, we can also extend this investigation to other state-of-the-art multilingual ASR systems such as Universal speech model (Zhang et al., 2023), ASR2K (Li et al., 2022), and DeepSpeech (Hannun et al., 2014) and on other tasks (*e.g.,* speaker verification (Toussaint and Ding, 2022)). Also, we only experimented with the original SMOTE (Chawla et al., 2002) framework, but improvements could be obtained with dedicated versions, *e.g.,* (Alex and Nayahi, 2023), (Dablain et al., 2023), (Maldonado et al., 2022). Our study focused on the Portuguese language but we are currently extending it to other languages. Finally, these results ask for a thorough analysis to detect the speech meta-features that trigger the disparate behavior of these ASR systems. For instance, correlational features among skin-tone scale and voice-timber in speech utterances affect the disparity gap in performance.

6 Conclusion

In this work, we presented an extensive study of recent ASR systems, namely, Whisper and MMS, in the light of stereotypical biases such as gender, age, skin tone, and geo-location, for the Portuguese language. Despite observing mild performance disparities concerning individual categories such as 'Age', 'Skin Tone Color', and 'Geo-location', we empirically show significant performance differences between the 'Male' and 'Female' subgroups. The first observation was to notice the imbalance in the various distributions, and that a naïve oversampling may further contribute to disparate performance behavior. This motivated us to employ SMOTE, and our results attested that oversampling technique has an overall beneficial impact in reducing performance differences. We also discuss some limitations of our study along with future work.

References

Martine Adda-Decker and Lori Lamel. 2005. Do speech recognizers prefer female speakers? In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, pages 2205–2208. ISCA.

Tao Xu Greg Brockman Christine Mcleavey Ilya Sutskever Alec Radford, Jong Wook Kim. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Suja A. Alex and J. Jesu Vedha Nayahi. 2023. Classification of imbalanced data using SMOTE and autoencoder based deep convolutional neural network. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, 31(3):437–469.

Caerwyn Ash, Godfrey Town, Peter Bjerring, and Samuel Webster. 2015. Evaluation of a novel skin tone meter and the correlation between fitzpatrick skin type and skin color. *Photonics & Lasers in Medicine*, 4:177 – 186.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Srinivas Bangalore, Dilek Z. Hakkani-Tür, and Gökhan Tür. 2005. Introduction to the special issue on spoken language understanding in conversational systems. *Speech Communication*, 48:233–238.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna M. Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5454–5476. Association for Computational Linguistics.

Jie Cao, Ananya Ganesh, Jon Cai, Rosy Southwell, E. Margaret Perkoff, Michael Regan, Katharina Kann, James H. Martin, Martha Palmer, and Sidney D'Mello. 2023. A comparative analysis of automatic speech recognition errors in small group classroom discourse. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23, page 250–262, New York, NY, USA. Association for Computing Machinery.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Damien Dablain, Bartosz Krawczyk, and Nitesh V. Chawla. 2023. Deepsmote: Fusing deep learning and SMOTE for imbalanced data. *IEEE Trans. Neural Networks Learn. Syst.*, 34(9):6390–6404.

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech Language*, 84:101567.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *ArXiv*, abs/2103.15122.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2019. 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, AI4TV@MM 2019, Nice, France, October 21, 2019*, pages 3–9. ACM.

Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Gregory Frederick Diamos, Erich Elsen, Ryan J. Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and A. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition. *ArXiv*, abs/1412.5567.

Drew Harwell. 2018. The Accent Gap.

Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Cantón Ferrer. 2021. Towards measuring fairness in ai: The casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4:324–332.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Atharva Kulkarni, Ajinkya Kulkarni, Miguel Couceiro, and Hanan Aldarmaki. 2023. Adapting the adapters for code-switching in multilingual asr.

X Li, F Metze, D. R. Mortensen, A. W. Black, and S Watanabe. 2022. Asr2k: Speech recognition for around 2000 languages without audio. *ArXiv*, abs/2209.02842.

Lanna Lima, Vasco Furtado, Elizabeth Furtado, and Virgilio de Almeida. 2019. Empirical analysis of bias in voice-based personal assistants. In *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 533–538. ACM.

Sebastián Maldonado, Carla Vairetti, Alberto Fernández, and Francisco Herrera. 2022. FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification. *Pattern Recognit.*, 124:108511.

Josh Meyer, Lindy Rauchenstein, Joshua D. Eisenberg, and Nicholas Howell. 2020. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *International Conference on Language Resources and Evaluation*.

David Molina, Leonardo Causa, and Juan E. Tapia. 2020. Reduction of bias for gender and ethnicity from face images using automated skin tone classification. *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5.

Ellis Monk. 2019. Monk skin tone scale.

Martin Mühleisen. 2018. The long and short of the digital revolution. *Finance Development*, 0055(002):A002.

Tanvina B. Patel and Odette Scharenborg. 2023. Using data augmentations and vtln to reduce bias in dutch end-to-end speech recognition systems. *ArXiv*, abs/2307.02009.

Bilal Porgali, Vítor Albiero, Jordan Ryda, Cristian Canton Ferrer, and Caner Hazirbas. 2023. The casual conversations v2 dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 10–17.

V Pratap, A Tjandra, B Shi, P Tomasello, A Babu, S Kundu, A Mamdouh E, Z Ni, A Vyas, M. Fazel-Zarandi, A Baevski, Y Adi, X Zhang, Wei-Ning Hsu, A Conneau, and M Auli. 2023. Scaling speech technology to 1, 000+ languages. *ArXiv*, abs/2305.13516.

M Sawalha and M Abu Shariah. 2013. 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 *2nd Workshop of Arabic Corpus Linguistics WACL-2*.

Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, EthNLP@EACL*, pages 53–59.

Rachael Tatman and Conner Kasten. 2017. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 934–938. ISCA.

Wiebke Toussaint and Aaron Yi Ding. 2022. Bias in automated speaker recognition. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Y Zhang, W. H., James Qin, Y Wang, A Bapna, Z Chen, N Chen, Bo Li, V Axelrod, G Wang, Z Meng, Ke Hu, A Rosenberg, R Prabhavalkar, D. S. Park, P Haghani, J Riesa, G Perng, H Soltau, T Strohman, B Ramabhadran, T. N. Sainath, Pedro J. Moreno, C-C Chiu, J Schalkwyk, F Beaufays, and Y Wu. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *ArXiv*, abs/2303.01037.