Factorizing Gender Bias in Automatic Speech Recognition for Mexican Spanish

Anastasiia Chizhikova, Hannah Billinghurst, Michelle Elizabeth, Shehenaz Hossain, Ajinkya Kulkarni, Gaël Guibon, Miguel Couceiro

Factorizing Gender Bias in Automatic Speech Recognition for Mexican Spanish

Anastasiia Chizhikova[1*], *Hannah Billinghurst*[1,*], *Michelle Elizabeth*[1,*], *Shehenaz Hossain*[1,*], *Ajinkya Kulkarni*[2], *Gaël Guibon*[3], *Miguel Couceiro*[3,4]

[1]IDMC, University of Lorraine, France,[2]IDIAP, Switzerland
[3]Université de Lorraine, CNRS, LORIA, Nancy, France
[4]INESC-ID, IST, Universidade de Lisboa, Portugal

Abstract

Advances in speech technologies have led to significant progress in large acoustic models such as Whisper and Multilingual Massive Speech (MMS), improving tasks like Automatic Speech Recognition (ASR). Yet, there is still a need for thorough research to recognize and tackle stereotypical biases. In this paper, we investigate Whisper and MMS systems to quantify gender bias and factorize gender bias considering voice timbre, skin tone, and age group for Mexican-Spanish in a multilingual ASR setting. In addition to traditional ASR evaluation such as word error rate and phoneme error rate, we also perform statistical significance tests. Furthermore, we explore the vital role of factorization of gender attributes into sub-groups in bias quantification. This work presents an initial study of gender inclusivity with various factors in the context of MMS and Whisper for Mexican-Spanish.

Index Terms: automatic speech recognition, bias, gender, Mexican Spanish

1. Introduction

The assessment by the World Economic Forum suggests that it would require 131 years at the current pace to narrow the global gender gap in economic participation and opportunity[1]. The advancements in Artificial Intelligence (AI) serve no purpose if they fail to reach the people who need them the most. It is imperative to avail universal access and develop AI tools with unbiased behavior even for low-resource languages and dialects. Recently, conversational AI, Natural Language Processing (NLP), and speech processing systems have become ubiquitous in our daily lives. Their purpose is to enhance universal access, elevate the quality of life, and offer essential services such as home assistance, question-answering systems, and automated call centers, among others. Moreover, it is vital to investigate AI tools and their impact on inclusivity and fair access.

Mexico is the largest Spanish-speaking country in the world, home to 113 million native speakers[2]. An acoustic comparative study of Spanish among speakers from Mexico and Spain illustrated differences in pronunciation of segments, syllabic duration, intensity, and frequency range [1]. There are numerous speech resources accessible for the Spanish language, yet those specifically focused on Mexican Spanish are relatively scarce [2]. Despite the availability of Mexican Spanish speech resources, researchers have only developed a limited database, requiring individuals to reach out directly to the authors to obtain access to these resources [3, 4, 5]. Despite this, many multilingual Automatic Speech Recognition (ASR) systems such as Whisper [6], Multilingual Massive Speech (MMS) [7] system, ASR2K [8], and Universal speech models [9] provide speech-to-text support to generalize Spanish without explicitly trained systems for Mexican Spanish. Hence, there are no specific versions of these state-of-the-art systems specifically created for Mexican Spanish.

Speech-to-text systems may incorporate biases of different types, and one of the most important ones is gender bias. Over time, various studies have explored different facets of gender within ASR systems. Many of these investigations concentrated on distinct groups within gender, age, and accent categories, assessing performance using metrics such as word error rate, phoneme error rate, and p-value [10, 11, 12, 13, 14, 15, 16]. In the case of the Dutch language, a bias analysis utilizing the Hidden Markov Model-deep neural network (HMM-DNN) ASR system to examine gender bias has been conducted [17, 18]. Following this, they suggested vocal tract length normalization and data augmentation methods as means to alleviate biases observed within gender and age demographics [19]. A similar study [20] examining bias in ASR systems for Portuguese revealed that incorporating gender alongside skin tone as a meta-attribute exposes significant disparities that might otherwise go unnoticed by solely focusing on gender differences.

In 2023, Meta AI published a Massively Multilingual Speech (MMS) model [7] supporting 1000+ languages for ASR along with a gender bias study. This gender bias study was conducted on a development set of the FLEURS dataset [21] over 27 languages including Spanish. However, no specific study has taken into account Mexican Spanish. To our knowledge, this is the only study conducted to analyze gender bias in Mexican Spanish. Therefore, there exists a comprehensive research gap in analyzing SOTA systems from a fairness-centric perspective.

In this paper, we present an empirical study of MMS and Whisper ASR systems by factorizing gender into various attributes such as age groups, skin tone, and voice timbre. The usage of additional meta-attributes allows the impact of other attributes in the quantification of bias. Furthermore, such factorization of bias attributes allows us to comprehend the contributing demographic factors. In comparison to [20], we also take into account the newly released Whisper variant Large-v3 in our work with voice timbre as additional attribute. We used Casual Conversation dataset version 2 [22] which is a multimodal multilingual fairness-centric dataset by Meta AI.

In [1] it is stated that Mexican speakers use a high-pitched speech with higher variations of the frequency range. This indicates that the usage of voice timbre can provide insight into biases aligned with gender against the ASR systems. For in-

*These authors contributed equally to this work

[1]https://www3.weforum.org/docs/WEF_GGGR_2023.pdf

[2]https://en.wikipedia.org/wiki/Mexican_Spanish

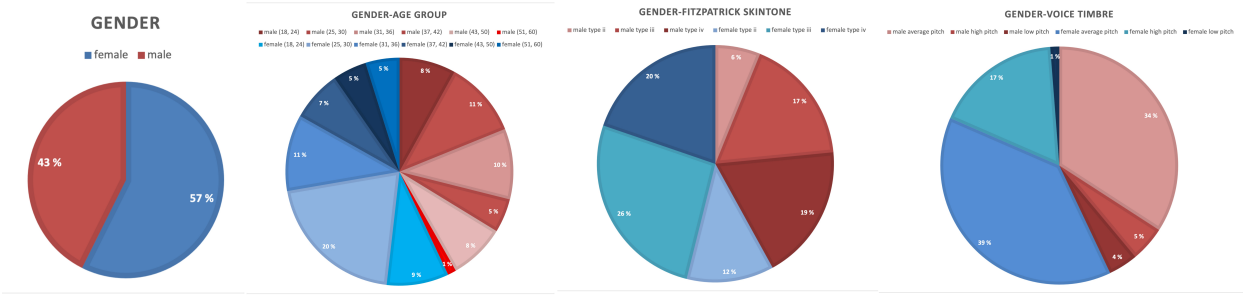

Figure 1: *Speech utterance distribution in CCv2 across gender, and gender with voice timbre, age groups, and Fitzpatrick skin tone type*

stance, analyzing gender bias alongside voice timbre encapsulates tonal information that is directly correlated with formants. Therefore, it provides vital information on performance degradation in cases where the female gender naturally has higher formant information compared to the male. In this study, we opted for traditional ASR evaluation metrics such as Word Error Rate (WER), and Phoneme Error Rate (PER) along with p-values as a statistical significance test. The main contributions of this paper are the following:

- The first study on disparities in Whisper and MMS ASR systems for Mexican Spanish on conversational speech.
- To our knowledge, it is the first empirical study to factorize gender bias with respect to attributes such as age, voice timbre and skin-tone.

Our results show that the different variants of Whisper present similar performances for both male and female voices. More precisely, we observed a slight bias towards male voices for all Whisper variants performed better. Nonetheless, from the p-value statistical tests, these disparities cannot be considered significant.

However, these disparities were drastically increased when factorizing gender w.r.t. skin tone and age groups. In particular, we observed that Whisper Large-v3, which achieves benchmarking performances on ASR datasets, presents the most disparate performance results w.r.t. skin tone. This fact was also observed when factorizing gender w.r.t. to voice timber.

Overall, it was surprising to observe behavior differences between different variants of Whisper Large, since they only differ in the training data used. This shows the importance of choosing training data that promotes inclusiveness.

2. Dataset Description

The Casual Conversations version 2 (CCv2) dataset is open-source and can be accessed through the Meta AI website[3] [22, 23]. 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. 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 [24] and Fitzpatrick Skin Type [25, 26]), voice timbre, the speaker's activity, categorized as gesture, action, or appearance. We opted, therefore, to avoid skewed comparison between skin-tone scales using Monk skin tone and only conducted a study using the Fitzpatrick skin type.

The CCv2 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 Mexican Spanish. The uniform scripted recordings with consistent textual content and phonetic variations facilitate analysis of meta-attributes influencing performance disparities. For Mexican Spanish, there were 253 speech utterances with an average duration of 90 seconds per utterance, totaling approximately 6 hours 18 mins. We illustrate the data distribution across gender and gender intersection with Fitzpatrick's skin tone, age group, and voice timbre in Figure 1. In the context of assessing the fairness of ASR systems, we focused on four annotated labels: gender, age groups, Fitzpatrick scale, and voice timbre. 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+.

3. Experimental Protocol

In this section, we describe the experimental setup for using Whisper and MMS ASR systems to measure disparities using the CCv2 dataset. Furthermore, we delve into the evaluation methodology opted in this work.

3.1. ASR Systems

Whisper: Whisper is a robust speech recognition model introduced by OpenAI in 2022. It leverages multitask learning on 680,000 hours of labeled multilingual recordings sourced from the Internet. These recordings include filtered transcriptions, covering approximately 96 languages across 117,000 hours of audio data. Incorporating the Transformer encoder-decoder architecture with multitask learning techniques, Whisper facilitates language identification, multilingual speech transcription, and word-level timestamps. By splitting input audio into thirty-second chunks, Whisper enhances the effectiveness of transcribing long recordings. Various Whisper variants are available, differing in model parameter sizes: Tiny (39 Million), Base (74 Million), Small (244 Million), Medium (769 Million), Large (1550 Million), and Large-v2 and v3 (1550 Million each). These models are categorized into English-only and multilingual variants. This paper focuses on investigating the Large-v1, Large-v2, and Large-v3 variants of Whisper.

Massively Multilingual Speech: In 2023, Meta AI launched the Massively Multilingual Speech (MMS) project, signifi-

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

cantly expanding language support to encompass over 1000 languages across various speech processing applications. The core elements of the MMS system consist of a unique dataset sourced from publicly available religious texts and proficient use of cross-lingual self-supervised learning techniques. This project covers a wide array of tasks including speech recognition, language identification, and speech synthesis. Built upon the Wav2Vec 2.0 architecture, MMS undergoes training via the integration of cross-lingual self-supervised learning and supervised pre-training for ASR. It incorporates language adapters that allow dynamic loading and interchange during inference, featuring multiple Transformer blocks, each enhanced with a language-specific adapter. The MMS system offers two variants based on model parameters, with 317 million and 965 million parameters. For this investigation, the MMS system with 965 million model parameters was utilized.

3.2. Evaluation Strategy

In assessing ASR systems, we employ the Word Error Rate (WER), a standard metric indicating the percentage of incorrectly recognized words. This approach enables objective comparison and identification of performance biases among the four ASR systems. Evaluation metrics including WER, Character Error Rate (CER), and Phoneme Error Rate (PER) are computed using the jiwer library[4].

We measure the statistical significance level of the score differences by calculating the p-value[5] [27] for those categories that only include two subcategories (gender) and the one-way ANOVA test when there are more than two subcategories in the experiment (namely, in the experiments with mixed categories). To tackle the problem of subcategories imbalance in the test data, which might lead to inadequate evaluation results, we discard those subcategories that do not contain enough samples to make any valid conclusions. Thus, for the gender category, we only consider *male* and *female* (i.e. always cis-female and cis-male in this study), whereas for the skin tones, we only consider Fitzpatrick types ii to iv.

4. Empirical Study

This section provides a comprehensive examination of ASR performance concerning gender, including gender factorization with attributes like skin tone, age groups, and voice timbre. From Table 1, bias analysis for ASR systems reveals notable patterns across various demographic factors. The bar plots for only the gender attribute is illustrated in Figure 2. We can see that in terms of gender, Whisper Large-v1 and v2 exhibit superior performance in WER for the two genders considered in this study. However, there seems to be a consistent bias favoring male voices over female voices across all models tested. Notably, Whisper-large-v3 demonstrates the least disparity in genders compared to other models, although the difference remains slight.

We then considered gender in conjunction with skin tone, which revealed disparate performances of the different Whisper variants. For instance, different skin tones are associated with varying error rates, with type 2 being optimal for female voices and type 3 for male voices. Furthermore, while MMS and Whisper Large-v2 display the least discrimination toward gender skin tones if we refer to the standard deviations, Whisper Large-v3 is the most discriminatory, aligning closely with its overall performance trends. On that note, Male and female skin-tones that are most likely discriminated are not the same: type 3 for female and type 2 for males. Even though Whisper Large-v3 has shown benchmarking performance on standard ASR datasets, for Mexican Spanish, it has shown the least inclusivity considering skin tone, as shown by the higher standard deviation (6.42) of this model compared to the other ones.

The analysis extends to gender-age-group factorization,

Feature\Model	MMS	W-v1	W-v2	W-v3
Gender				
female	13.89	13.15	**12.75**	18.55
male	11.15	**09.98**	10.53	16.35
model stdev	2.74	3.17	2.22	2.20
Gender-Skin-Tones				
female-t2	08.03	16.54	**07.19**	15.23
female-t3	16.60	**07.16**	15.46	21.44
female-t4	13.95	**12.40**	12.64	16.92
female-t5	10.74	08.72	**08.05**	08.72
male-t2	17.99	**16.69**	16.71	26.73
male-t3	07.53	**06.22**	06.52	11.74
male-t4	12.17	**11.25**	12.40	16.56
male-t5	14.18	12.00	**10.49**	26.01
model stdev	3.77	3.92	3.79	6.42
Gender-Age-Group				
female-18-24	12.43	10.62	**10.43**	15.24
female-25-30	12.15	12.58	**11.08**	17.80
female-31-36	11.29	**09.81**	10.15	15.14
female-37-42	16.37	15.12	**14.84**	22.13
female-43-50	12.47	**11.07**	11.58	16.41
female-51-60	14.18	13.87	**13.48**	17.48
male-18-24	10.86	09.65	**09.55**	14.66
male-25-30	14.33	**12.73**	14.29	20.72
male-31-36	10.23	**09.46**	11.26	12.74
male-37-42	08.84	**08.00**	08.42	18.93
male-43-50	08.14	06.93	**05.27**	13.32
male-51-60	21.48	**19.91**	20.69	28.19
model stdev	3.61	3.52	3.84	4.34
Gender-voice timbre				
female-high	12.12	**10.93**	11.15	16.71
female-avg	13.32	12.83	**11.93**	17.76
female-low	04.03	**03.47**	04.03	11.19
male-high	05.84	**04.39**	06.06	12.11
male-avg	12.63	**11.05**	11.87	17.46
male-low	04.97	**03.76**	03.86	12.01
model stdev	4.30	4.30	3.92	3.07

Table 1: *Word Error Rates for each feature and standard deviation for each model per bias type. Lower is better. The best scores are in bold, and the second best is underlined. W refers to Whisper-large versions.*

ASR systems	MMS	W-v1	W-v2	W-v3
p-value	0.5522	0.513	0.7008	0.7139

Table 2: *p-values for all ASR systems. W-vx refers to Whisper-large versions.*

[4] https://pypi.org/project/jiwer/

[5] Due to the page limit, we only provide p-values for gender attributes, and p-values for other groups can be found in the supplementary material.

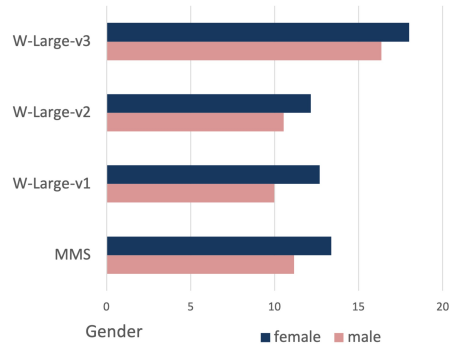

Figure 2: *Bar plots depicting Whisper and MMS ASR performances for gender attribute*

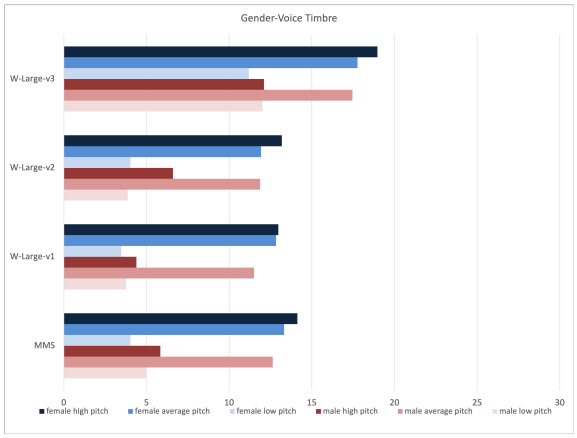

Figure 3: *Bar plots illustrating Whisper and MMS ASR performances for gender and voice timbre attributes together*

where Whisper Large-v2 and v1 demonstrate the best performance across age groups. Conversely, Whisper Large-v3 consistently exhibits the poorest performance in terms of WER. Whisper Large-v1 seems to be the least discriminant towards age groups, along with MMS. The examination of gender in voice timbre reveals noteworthy trends as shown in Figure 3. Female voices with higher pitch exhibit higher average error rates compared to male voices, suggesting a bias towards certain voice characteristics. The most difficult voice timbres are the same for the two genders: average pitch seems more difficult to handle across all ASRs. Prominently, we observed that ASRs performed well on male high voice timbre in opposition to female high voice timbre, which yields a high word error rate. This provides insight into the effect and differentiation in voice characteristics causing an adverse impact on ASR performance. It is important to mention that Whisper Large-v3 undergoes training with a more extensive multilingual dataset compared to other variants. The p-value analysis for ASR performances on Word Error Rate (WER) provides insights into the statistical significance of observed differences among the ASR systems. A p-value below a predetermined threshold, typically 0.05, indicates statistical significance, suggesting that the observed differences in WER between ASR systems are unlikely to have occurred by chance. In the presented Table 2, the p-values for MMS, Whisper Large-v1, Whisper Large-v2, and Whisper Large-v3 are 0.5522, 0.513, 0.7008, and 0.7139, respectively. These values suggest that there is no significant difference in WER between the ASR systems, as none of the p-values fall below the threshold. Therefore, based on the p-value analysis, we cannot conclude that one ASR system performs significantly better or worse than another in terms of WER. Therefore, analyzing biases by considering a single attribute such as gender might not present other aligned latent attributes such as skin tone, age-group, and voice timbre.

5. Discussion and Conclusion

In this work, we present a thorough investigation of the performance of some recently proposed ASR systems with state-of-the-art performances, namely, the MMS and Whisper variants for Mexican Spanish. We conducted a comprehensive examination of ASR performance with respect to gender, and its intersectionality with attributes like skin tone, age groups, and voice timbre, revealing significant insights into biases present in ASR systems. While certain models such as Whisper Large-v1 and v2 demonstrate superior performance in gender identification, a consistent potential bias favoring male voices over female voices persists across all models tested.

Notably, Whisper Large-v3 shows the least disparity when considering gender-only attributes. The Whisper Large-v3 has shown better WER on common-voice 15 and FLEURS evaluation sets[6]. However, Whisper Large-v3 exhibits potential biases in consideration of gender factorized with age groups and voice timbre. The examination of gender in voice timbre uncovers biases favoring certain voice characteristics, with higher pitch female voices exhibiting higher error rates compared to male voices. We observed varying performance from ASR systems across skin tone with gender, this suggests the need for further investigation. Moreover, the p-value analysis conducted on ASR performances, specifically focusing on the Word Error Rate (WER), reveals that there are no significant differences in performance among the tested ASR systems. This finding suggests that solely analyzing biases based on a single attribute like gender may not comprehensively account for other related latent attributes such as skin tone, age group, and voice timbre. All ASR systems examined in this study generalize Spanish without distinguishing between Latin-American and European Spanish. This emphasizes the necessity for the development of ASR systems specifically tailored for Mexican Spanish, utilizing state-of-the-art architecture. Overall, these analyses shed light on the complex interplay of demographic factors in ASR performance and underscore the importance of addressing biases to ensure equitable outcomes across diverse user groups.

6. Limitation

We explored gender bias in conversational speech scenarios for Mexican Spanish. However, there's a pressing need to develop an evaluation dataset tailored to Mexican Spanish, encompassing reading speech settings similar to the Artie-Bias dataset [28] for English. Furthermore, we can expand this empirical inquiry to encompass other multilingual ASR systems, like the Universal Speech Model [9] and ASR2K [8], and to diverse tasks such as speaker verification and deepfake detection. As a preliminary approach towards bias mitigation, we can employ data augmentation techniques designed for distinct categories in the future. While analyzing biases using a single attribute may not reveal causal factors, incorporating various meta-attributes like voice timbre and leveraging explainable AI can facilitate the development of inclusive ASR systems.

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

7. References

[1] E. P. V. Patiño, "Prosodic comparative study of Mexico city and Madrid Spanish," *Speech Prosody*, 2008.

[2] C. D. H. Mena, I. V. M. Ruiz, and J. A. H. Camacho, "Automatic speech recognizers for Mexican Spanish and its open resources," *Journal of Applied Research and Technology*, vol. 15, 2019.

[3] E. Uraga and C. Gamboa, "VOXMEX speech database: Design of a phonetically balanced corpus," in *International Conference on Language Resources and Evaluation*, 2004.

[4] I. Kirschning, "Research and development of speech technology & applications for Mexican Spanish at the Tlatoa group," *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, 2001.

[5] J. M. O. Espinoza, P. Mayorga-Ortiz, H. Hidalgo-Silva, L. Vizcarra-Corral, and M.-L. Mendiola-Cárdenas, "VoCMex: a voice corpus in mexican spanish for research in speaker recognition," *International Journal of Speech Technology*, vol. 16, pp. 295 – 302, 2012.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202, 2023, pp. 28 492–28 518.

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

[8] X. Li, F. Metze, D. R. Mortensen, A. W. Black, and S. Watanabe, "ASR2K: speech recognition for around 2000 languages without audio," in *Proc. of INTERSPEECH*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 4885–4889.

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

[10] M. Adda-Decker and L. Lamel, "Do speech recognizers prefer female speakers?" in *Proc. of INTERSPEECH*. ISCA, 2005, pp. 2205–2208.

[11] M. Garnerin, S. Rossato, and L. Besacier, "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*, R. Troncy, J. Laaksonen, H. R. Tavakoli, L. J. B. Nixon, and V. Mezaris, Eds. ACM, 2019, pp. 3–9.

[12] M. Sawalha and M. A. Shariah, "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*, 2013.

[13] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[14] R. Tatman, "Gender and Dialect Bias in Youtube's Automatic Captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, EthNLP@EACL*, 2017, pp. 53–59.

[15] R. Tatman and C. Kasten, "Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and Youtube Automatic Captions," in *Proc. of INTERSPEECH*, F. Lacerda, Ed. ISCA, 2017, pp. 934–938.

[16] D. Harwell, "The Accent Gap," 2018. [Online]. Available: https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/

[17] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying Bias in Automatic Speech Recognition," *ArXiv*, vol. abs/2103.15122, 2021.

[18] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer speech and Science*, vol. 84, p. 101567, 2024.

[19] T. B. Patel and O. Scharenborg, "Using Data Augmentations and VTLN to Reduce Bias in Dutch End-to-End Speech Recognition Systems," *ArXiv*, vol. abs/2307.02009, 2023.

[20] A. Kulkarni, A. Tokareva, R. Qureshi, and M. Couceiro, "The Balancing Act: Unmasking and Alleviating ASR Biases in Portuguese," *ArXiv*, vol. abs/2402.07513, 2024.

[21] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "FLEURS: Few-shot learning evaluation of universal representations of speech," *IEEE Spoken Language Technology Workshop (SLT)*, 2022.

[22] B. Porgali, V. Albiero, J. Ryda, C. C. Ferrer, and C. Hazirbas, "The Casual Conversations v2 dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2023, pp. 10–17.

[23] C. Hazirbas, J. Bitton, B. Dolhansky, J. Pan, A. Gordo, and C. C. Ferrer, "Towards measuring fairness in AI: The casual conversations dataset," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, pp. 324–332, 2021.

[24] C. M. Heldreth, E. P. Monk, A. T. Clark, C. Schumann, X. Eyee, and S. Ricco, "Which skin tone measures are the most inclusive? an investigation of skin tone measures for Artificial Intelligence." *ACM Journal on Responsible Computing*, 2023.

[25] D. Molina, L. Causa, and J. E. Tapia, "Reduction of Bias for Gender and Ethnicity from face images using automated skin tone classification," *International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–5, 2020.

[26] C. Ash, G. Town, P. Bjerring, and S. Webster, "Evaluation of a novel skin tone meter and the correlation between Fitzpatrick skin type and skin color," *Photonics & Lasers in Medicine*, vol. 4, pp. 177 – 186, 2015.

[27] P. Mishra, U. Singh, C. M. Pandey, P. Mishra, and G. Pandey, "Application of student's t-test, analysis of variance, and covariance," *Annals of Cardiac Anaesthesia*, vol. 22, pp. 407 – 411, 2019.

[28] J. Meyer, L. Rauchenstein, J. D. Eisenberg, and N. Howell, "Artie Bias corpus: An open dataset for detecting demographic bias in speech applications," in *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. European Language Resources Association, 2020, pp. 6462–6468.

1. Phoneme error rates

In this section we report on the phoneme error rates (PER) for all models on the CCv2 dataset.

Feature/Model	MMS	W-v1	W-v2	W-v3
Gender				
female	08.14	**08.07**	**08.07**	12.88
male	06.38	**05.93**	06.59	11.87
Gender-Skin-Tones				
female-t2	01.75	**01.50**	02.00	08.38
female-t3	10.42	10.87	**10.21**	15.43
female-t4	09.00	**08.41**	08.97	12.38
male-t2	12.04	**11.65**	11.85	21.01
male-t3	03.53	**02.96**	03.50	08.02
male-t4	06.99	**06.71**	07.75	11.81
Gender-Age-Group				
female-18-24	06.55	**06.18**	06.43	10.19
female-25-30	07.61	08.31	**07.60**	13.41
female-31-36	06.97	**06.27**	06.77	11.25
female-37-42	12.26	11.67	**11.66**	17.4
female-43-50	07.98	**07.61**	08.51	12.23
female-51-60	09.90	**09.67**	10.21	13.14
male-18-24	05.84	**05.27**	05.67	09.77
male-25-30	08.57	**07.91**	09.30	15.55
male-31-36	05.49	**05.17**	06.87	08.59
male-37-42	06.33	**05.60**	05.67	14.78
male-43-50	03.34	03.25	**02.30**	9.44
male-51-60	18.19	18.23	**18.17**	24.64
Gender-Voice-Timbre				
female-high	07.71	**07.27**	07.73	12.3
female-avg	08.54	08.64	**08.41**	13.27
female-low	**01.34**	01.44	02.21	08.86
male-high	**02.19**	**01.53**	02.95	08.08
male-avg	07.53	**07.12**	07.75	12.85
male-low	01.60	01.13	**01.10**	08.11

Table 1: *Phoneme Error Rates for each feature. Lower is better. Best scores are in bold, second best are underlined. W refers to Whisper-large versions.*

2. Further statistical tests

Below we report on the p-values for the different gender factorizations w.r.t age, skin tone and voice timber.

Feature/Model	MMS	W-v1	W-v2	W-v3
Gender	0.5522	0.5130	0.7008	0.7139
Gender-Skin-Tone	0.4650	0.3947	0.5391	0.2929
Gender-Age-Group	0.9720	0.9540	0.9412	0.8395
Gender-Voice-Timbre	0.5519	0.5210	0.6808	0.8309

Table 2: *p-values for each feature. p-value less than 0.05 are indicated with a *. W refers to Whisper-large versions.*