Unveiling Biases while Embracing Sustainability: Assessing the Dual Challenges of Automatic Speech Recognition Systems

Ajinkya Kulkarni, Atharva Kulkarni, Isabel Trancoso, Miguel Couceiro

Unveiling Biases while Embracing Sustainability: Assessing the Dual Challenges of Automatic Speech Recognition Systems

Ajinkya Kulkarni[1,2], *Atharva Kulkarni*[3], *Miguel Couceiro*[4,5], *Isabel Trancoso*[5]

[1]IDIAP, Switzerland, [2]MBZUAI, UAE, [3]Erisha Labs, India
[4]Université de Lorraine, CNRS, LORIA, Nancy, France
[5]INESC-ID, IST, Universidade de Lisboa, Portugal

ajinkya.kulkarni@idiap.ch

Abstract

In this paper, we present a bias and sustainability focused investigation of Automatic Speech Recognition (ASR) systems, namely Whisper and Massively Multilingual Speech (MMS), which have achieved state-of-the-art (SOTA) performances. Despite their improved performance in controlled settings, there remains a critical gap in understanding their efficacy and equity in real-world scenarios. We analyze ASR biases w.r.t. gender, accent, and age group, as well as their effect on downstream tasks. In addition, we examine the environmental impact of ASR systems, scrutinizing the use of large acoustic models on carbon emission and energy consumption. We also provide insights into our empirical analyses, offering a valuable contribution to the claims surrounding bias and sustainability in ASR systems.

Index Terms: ASR, Bias, carbon footprint, sustainability

1. Introduction

The advent of large deep neural networks (DNNs) has brought about substantial advancements in various speech-processing applications, notably in speech recognition. However, amidst this progress, there remains a notable gap in understanding the inherent biases towards gender, age groups, and accents. For instance, home assistant devices often exhibit biased performances towards non-native English speakers [1], limiting access to technology for certain individuals, particularly in speech-enabled human-machine interfaces [2, 3]. This exclusionary behavior may impede the usability of crucial services, such as emergency assistance for the elderly or navigation aids for differently-abled individuals. Thus, comprehensive studies of these large DNN models are crucial to ensuring their widespread and inclusive use.

In recent years, there has been a growing research community dedicated to examining biases in automatic speech recognition (ASR) systems, more specifically for English [4, 5, 6]. This research primarily focuses on evaluating the disparities related to gender, age, accent, dialect, and racial attributes [7, 8, 9, 10, 11, 12, 13]. However, the training of large DNNs with extensive datasets necessitates ever-increasing computational resources, directly contributing to carbon emissions [14]. This environmental impact extends even to the inference time of these large DNN systems. The substantial release of CO2 into the atmosphere poses a significant threat to life on Earth, with consequences often overlooked within the deep learning research community, where effective benchmarking is lacking. These emissions not only disrupt the delicate balance of ecosystems but also raise profound ethical concerns regarding our responsibility towards the environment. Hence, investigating the carbon footprint and energy consumption of deep learning models is imperative for the sustainable development of deep learning systems. In 2018, the Intergovernmental Panel on Climate Change emphasized the importance of limiting global temperature rise to below 1.5°C to mitigate adverse impacts on various aspects such as extreme weather events, ecosystems, and carbon removal efforts [1]. To address this concern, efforts have been made to estimate carbon emissions and energy consumption. The Experimental Impact Tracker library, published in 2019 [15], was one such initiative. Subsequently, other carbon footprint tracking platforms like codecarbon[2], carbontracker[3], and eco2ai [16] have emerged, providing support for tracking activity across GPU, CPU, user interfaces, and seamless integration into Python scripts. These tools enable organizations to compute carbon emissions and energy consumption, facilitating the development of sustainable AI systems.

The majority of studies on quantifying bias in ASR systems revolve around training individual systems and analyzing the disparities in ASR performance across different bias categories. For example, studies such as [17, 18] conducted bias analysis for Dutch using the Hidden Markov Model-DNN ASR system to observe gender bias. Subsequently, techniques such as vocal tract length normalization and data augmentation were proposed to mitigate biases in gender and age groups [19]. Literature also suggests that ASR systems can exhibit bias at various stages of development, including data curation, model architecture design, and evaluation protocols [20]. In 2018, an article titled "The Accent Gap" published in the Washington Post [1] illustrated inconsistent performance across accents on commercial home assistant systems for non-native English speakers. Similarly, gender and accent analyses on YouTube's captions were described in [21, 22], emphasizing the need for sociolinguistically-stratified validation of ASR systems. More recently, [23] established biases in multilingual ASR systems for Portuguese. Despite these efforts, there remains a lack of comprehensive studies focused on larger ASR systems for English.

ASR systems have progressively advanced to support over 100 languages [24, 25, 26, 27]. However, with the growing size of models and the vast amount of training data, computational requirements are also escalating, leading to increased carbon footprint and energy consumption. Research presented in [28] detailed the training cost of ASR systems in terms of energy and carbon footprint, alongside improved performance. However, it is also essential to measure the carbon footprint during inference, given the widespread use of ASRs in both industry and society. Once deployed in real-world scenarios, it becomes

[1]https://www.ipcc.ch/sr15/download/
[2]https://codecarbon.io/
[3]https://carbontracker.org/

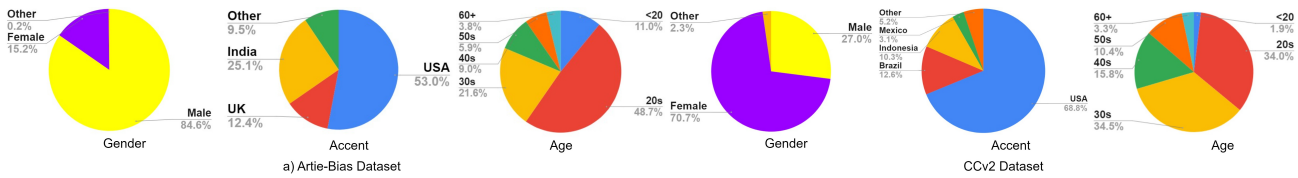

Figure 1: *Speech utterance distribution across gender, accent, and age for Artie-Bias and CCv2 dataset.*

challenging to backtrack, underscoring the importance of making users aware of their environmental impact and guiding them in selecting systems that align with their needs [29, 30].

This paper addresses the dual challenges within English language ASR systems, focusing on potential biases and carbon footprint. We examine two recently proposed ASR variants, namely, the MMS [26] and Whisper [25] ASR models. Our empirical study delves into two fairness-centric evaluation datasets, the Artie-bias dataset [31] and the Casual conversation dataset version 2 [32], which feature variations of read and spontaneous speech, respectively. To evaluate their energy consumption and carbon footprint, we employ three different carbon tracking libraries, across four distinct NVIDIA GPU systems. This approach allows for an integrated analysis of the impacts of using various GPU systems on different ASR models. The key contributions of this paper are as follows: **1.** We assess different variants of Whisper and MMS ASR systems on two bias-focused datasets, unveiling hidden disparities between read and spontaneous speech regarding gender, accent, and age. **2.** We present the first systematic sustainability study that not only compares different large ASR systems but also benchmark carbon tracking tools with various GPU variants.

Our findings reveal that for English, and easily replicable in other languages, *the Whisper variants perform better than MMS* on read speech for all three categories (accent, age, and gender). However, there is a *drastic performance degradation of Whisper in spontaneous speech*. For individual categories, we observed significant performance differences between the two types of models. Also, it was surprising to observe that *larger versions of Whisper underperformed compared to the medium version*. This was particularly evident on the age category, where *all models behaved consistently better for higher age groups*. As for the carbon footprint at inference, we observed clear disparities between the different carbon tracking tools. Although all show similar measurements for both datasets, *eco2ai consistently underestimated carbon emissions*, compared to both carbontracker and code carbon. Also, when refining by GPU and configuration, *wide GPU bandwith seems to have a positive impact in both carbon emissions and energy comsumption.*

2. Dataset Description

Two datasets were selected to investigate biases in terms of age, gender, and accent: the Artie-Bias on read speech, and the CCv2 targeting spontaneous speech.

2.1. Artie-Bias Dataset

The Artie Bias dataset [31] is based on the test set partition of the English Common Voice corpus, released in June 2019, totaling 1712 utterances (approx 2.4 hours), in reading mode. Demographic information for each speaker includes gender (3 categories), age (8 groups), and accent (17 English accents). We have detailed the speech utterance count distribution across gender, accent, and age in Fig 1.a. The corpus is not balanced in terms of gender, as male speakers account for 1,431 utterances, whereas female speakers total only 257. Due to the skewed representation of accents, we only consider accents from the United States (US), India, and United Kingdom (UK). Similarly, we only take into account 6 age group intervals varying from less than twenty (< 20) to sixty and above (≥ 60) in 10-year groups.

2.2. Casual conversation dataset version 2

In 2023, Meta AI published the casual conversation dataset version 2 (CCv2) [32, 33], a fairness-centric self-recorded multilingual videos from different demographic world regions, with 5567 unique speakers. The dataset includes various self-labeled attributes such as gender, age, accent, location, skin tone, voice timbre, etc. In this study, we focused investigation only on gender, age, and accent for the English language. We have illustrated the sample distribution for various attributes in Fig 1.b. on a total of 1053 speech utterances. In the CCv2 dataset, the textual content of all the speech utterances remains the same across speakers from all the demographics. This allows us to observe the impact of biases without considering the pertaining variations in performances due to textual content. For accent bias, we only considered accent variants from US English speakers, Indonesia, Brazil, and Mexico. Furthermore, we used the same age-group intervals as used in the Artie Bias dataset.

3. ASR Systems

To evaluate the widespread usage of ASRs in both societal and industrial applications, we focused on the MMS and Whisper ASR systems, which have become reference SOTA models.

3.1. Massively Multilingual Speech system

In 2023, Meta AI launched the Massively Multilingual Speech (MMS) project, as detailed in [26], which significantly broadened its language coverage to include more than 1100 languages across various speech processing applications. The MMS project encompasses a range of tasks, including speech recognition, language identification, and speech synthesis. Built upon the wav2vec 2.0 architecture [34], MMS has been trained using a combination of cross-lingual self-supervised learning and supervised pre-training for ASR. It integrates language adapters that can be dynamically loaded and swapped during inference, featuring multiple Transformer blocks, each enhanced with a language-specific adapter. The MMS system is offered in two variants based on model parameters, comprising 317 million and 965 million parameters. For this investigation, we employed the MMS system with 965 million model parameters.

3.2. Whisper

Whisper [25], introduced by OpenAI[4] in 2022, is a robust speech recognition model. It is trained using multitask learning on a dataset of 680k hours of labeled multilingual recordings,

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

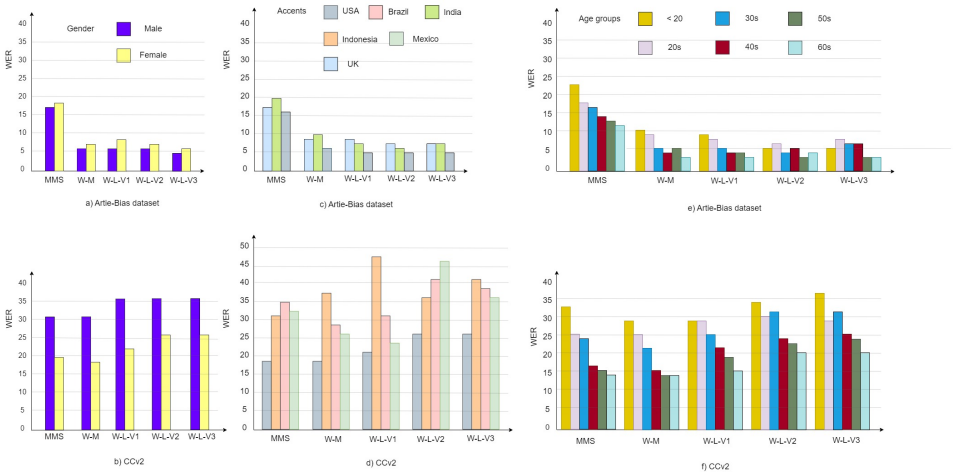

Figure 2: *Bar plots depicting Whisper and MMS ASR performances across gender, accent, and age. Whisper ASR variants are indicated respectively as Whisper-Medium (W-M), Whisper-Large (W-L), Whisper-Large-V2 (W-L-V2), and Whisper-Large-V3 (W-L-V3).*

Table 1: *The table provides p-values for the gender category w.r.t. all models, and both Artie-bias and CCv2 datasets, where a p-value of 0.05 or lower is statistically significant.*

Datasets	W-M	W-L-V1	W-L-V2	W-L-V3	MMS
Artie-Bias	0.479947	0.188101	0.238292	0.473756	0.241962
CCv2	1.60E-08	0.00148	0.000121	0.000414	4.26E-06

supporting 96 languages. Whisper primarily utilizes the Transformer encoder-decoder architecture, to provide a multi-task learning framework across applications such as ASR, language identification, speech translation, etc. Whisper models are primarily categorized into two groups based on languages and tasks: English-only and multilingual models. In this study, we explore English-only variants of Whisper, including Medium (769 Million), Large-v1 (1550 Million), Large-v2 (1550 Million), and Large-v3 (1550 Million). Noticeably, the variants of Large mainly differ in the amount of training data used.

4. Experimental Setup

This section presents the experimental setup to investigate bias and sustainability issues pertaining to the ASRs of Section 3.

4.1. Bias study

For the evaluation of both models, we use Word Error Rate (WER) for comparison purposes with the literature, and we also report on the p-values[5] for statistical significance. These were obtained by pairwise ANOVA test with a threshold of 0.05, following the same protocol from [31]. This allows us to compare results objectively and to identify performance biases in the 5 ASR systems. We utilized the jiwer[6] library to compute WER, CER, and PER metrics, with preprocessing involving the English text normalizer from Whisper.

[5]Due to the page limit, we only provide p-values for gender groups, and provide the p-values for accent and age groups in the supplementary material along with Character Error Rate (CER) and Phoneme Error Rate (PER).

[6]`https://pypi.org/project/jiwer/`

4.2. Sustainability study

We opted for 3 different platforms to measure the carbon emission intensity and energy consumption, namely, codecarbon, carbontracker, eco2ai [16]. We incorporated these tools during inference of ASR on 20 mins of speech utterances across 4 NVIDIA GPUs, namely, RTX-5000-16GB, RTX-A5000-24GB, A100-40GB, and A6000-48GB. All experiments were carried out utilizing a cloud service provider based in Tamil Nadu, India, employing 32GB of RAM and 7 CPU cores. We repeated the inference run 3 times to take into account variations and utilized average estimates of energy consumption and carbon emission. System energy consumption can be quantified in either Joules (J) or watt-hours (Wh), with the latter being a unit of energy equivalent to the sustained operation of one kilowatt of power for one hour. Our study concentrated on assessing the energy usage of the GPU, CPU, and RAM, given their direct influence on the ASR inference process. Emissions across countries differ due to climate, geography, economy, fuel usage, and technological advancements. To mitigate regional differences, tools utilize emission intensity coefficients, indicating CO2 emissions per megawatt-hour of electricity. Governments incorporate these coefficients into environmental policy assessments to address emissions disparities [35].

5. Empirical Study

In this section, we present a comprehensive analysis on dual challenges in ASR concerning biases and sustainability. We provide detailed analyses of disparate performances across categories including gender, accent, and age groups on both read and spontaneous speech. Then, we present the empirical findings on sustainability and their analysis.

5.1. Results: ASR Bias

We illustrate the ASR performances on gender category in Fig 2.a. and 2.b. for read and spontaneous speech respectively. From the bar plots, we clearly observe larger disparities in spontaneous speech across all systems, with lower performances for the male gender. Interestingly, this tendency is reversed on read speech, where the Whisper variants outperform the MMS. For spontaneous speech, female speech performed better than male

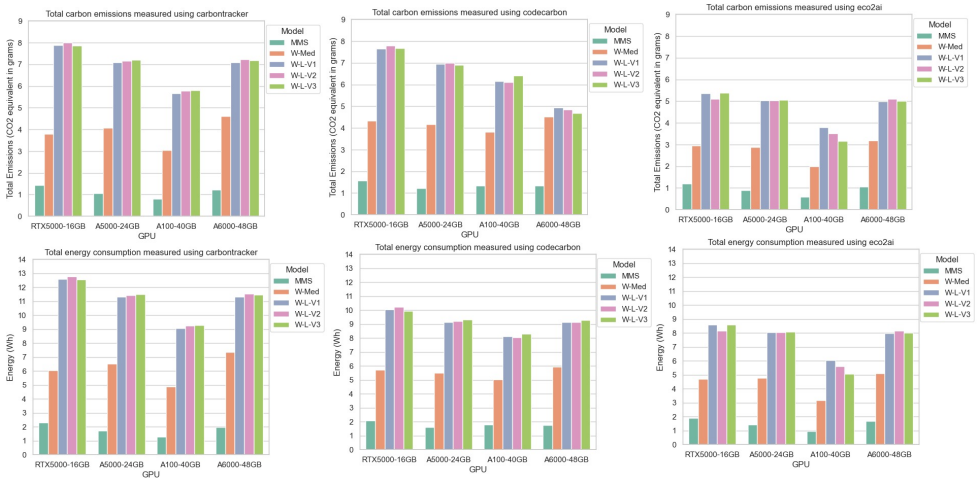

Figure 3: *Bar plots depicting carbon emissions (first row) and energy consumption (second row) of MMS and Whisper variants (W-M, W-L, W-L-V2, W-L-V3) obtained by carbontracker, codecarbon and eco2ai, as described in Subsection 4.2. The bar clusters correspond in each bar plot correspond to the 4 NVIDIA GPUs, namely, RTX-5000-16GB, RTX-A5000-24GB, A100-48GB, and A6000-48GB.*

and vice-versa for read speech, which is probably a reflection of the unbalanced training datasets across genders due to sociolinguistic differences [9, 10]. We observed continual improvement in Whisper large-v1 to v3 variants across both datasets, but Whisper medium performed better than other variants.

From Table 1, it is also noteworthy that gender differences are not significant for read speech, unlike spontaneous speech. From Fig 2.c and 2.d, we observe subtle differences in Whisper performances across accents in both datasets due to variations in model parameters. Overall, the US accent performed better than the others across both read and spontaneous speech. Noticeably, the MMS system shows higher WER in the case of read speech, while it performed better in spontaneous speech scenarios in the accent category. This illustrates the expected disparities between native and non-native English speakers. We illustrate performance disparities in ASR across the 6 age groups in Fig 2.e and 2.f for read and spontaneous speech, respectively. In the case of CCv2, the Whisper medium performed better than the other ASRs, irrespective of their model parameters. It is noteworthy that younger age groups, such as those under 20 and in their 20s, exhibit higher WERs than older age groups.

5.2. Results: Sustainability

We report our findings on both carbon emissions and energy consumption obtained using 3 commonly used platforms to assess the carbon footprint of deep learning models, namely, carbontracker, codecarbon and eco2ai, in the bar plots of Fig 3. We can observe a clear advantage of using MMS over the Whisper variants throughout all platforms. As expected, we see a similar behavior of the 3 Whisper Large variants, which is due to the fact that these only differ in the amount of training data used, and thus not impacting the inference step. Also, Whisper Medium shows better performances than the Whisper Large variants, clearly due the fact that the latter have twice the number of parameters of the former. As for the results produced by the 3 carbon tracking platforms (Fig 4), we can observe a slightly optimistic view provided by eco2ai. However, all platforms indicate similar trends for the 5 ASR systems considered. Perhaps more interesting is the slight advantage of NVIDIA GPU A100-40GB over the other NVIDIA GPUs. This could be explained by the fact that NVIDIA GPU A100-40GB has a much wider GPU bandwidth (1555 GB/s) than the other 3.

6. Discussion and Conclusion

We presented a thorough investigation of recently proposed ASR systems with state-of-the-art performances, namely, the MMS and Whisper variants, in both read (Arti Bias) and spontaneous (CCv2) settings, and from bias and sustainability perspectives. We also compared different NVIDIA GPU architectures for their carbon footprint and reported results obtained by 3 widely used carbon tracking platforms. Overall, as expected, ASR systems behave consistently better for read speech than for spontaneous speech, in terms of WER. Our findings indicate better bias performances (accent, age groups, and gender) for Whisper variants than for MMS, on read speech. However, this disparate behavior tends to disappear in spontaneous settings, where positive bias towards female speech is observed across all 5 ASRs considered. We also observed better ASR performances in older groups. Interestingly, larger variants of Whisper tend to behave worse than the medium counterpart.

Motivated by the widespread use of ASR systems, we also investigated carbon emissions and energy consumption in inference. Our results confirmed the superiority of MMS over the Whisper variants in terms of sustainability. It is noteworthy that MMS uses language-specific adapters, which restrict vocabulary output tokens, unlike the English-only variants of Whisper which have over 50K tokens. This distinction could potentially affect emission and energy consumption metrics. Indeed, language-specific adapters can help us save carbon emissions and mitigate biases. One should conduct a comprehensive analysis and perform a suitable study on each type of ASR usage. By the comparative study of the 4 NVIDIA GPUs, we observed the impact of bandwidth on carbon footprint, with NVIDIA GPU A100 showing better performances than the 3 other NVIDIA GPUs. It was also interesting to remark that eco2ai consistently reported lower carbon emissions and energy consumption than carbontracker and codecarbon. These findings emphasize the need for a comprehensive analysis of ASR systems that consider the diversity of performance metrics, implementations, and evaluation methodologies, *e.g.*, emission tracking and energy consumption.

7. Acknowledgements

The third named author was partially supported by the ANR project "Intrinsic and extrinsic evaluation of biases in LLMs" (InExtenso), ANR-23-IAS1-0004-01. The last named author was partially supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 and by Fundação para a Ciência e Tecnologia through the INESC-ID multi-annual funding from the PIDDAC programme (UIDB/50021/2020).

8. References

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

[2] Z. Mengesha, C. Heldreth, M. Lahav, J. Sublewski, and E. Tuennerman, ""i don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on african americans," *Frontiers in Artificial Intelligence*, 2021.

[3] K. Wenzel, N. Devireddy, C. Davison, and G. F. Kaufman, "Can voice assistants be microaggressors? cross-race psychological responses to failures of automatic speech recognition," *CHI Conference on Human Factors in Computing Systems*, 2023.

[4] J. Choe, Y. Chen, M. P. Y. Chan, A. Li, X. Gao, and N. R. Holliday, "Language-specific effects on automatic speech recognition errors for world englishes," in *International Conference on Computational Linguistics*, 2022.

[5] M. P. Y. Chan, J. Choe, A. Li, Y. Chen, X. Gao, and N. R. Holliday, "Training and typological bias in asr performance for world englishes," in *INTERSPEECH*, 2022.

[6] R. Sanabria, N. Bogoychev, N. Markl, A. Carmantini, O. Klejch, and P. Bell, "The edinburgh international accents of english corpus: Towards the democratization of english asr," *ICASSP*, 2023.

[7] L. Lima, V. Furtado, E. Furtado, and V. de Almeida, "Empirical Analysis of Bias in Voice-based Personal Assistants," in *Companion of The World Wide Web Conference*, 2019.

[8] S. L. Blodgett, S. Barocas, H. D. III, and H. M. Wallach, "Language (Technology) is Power: A Critical Survey of "Bias" in NLP," in *ACL*, 2020.

[9] 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*, 2020.

[10] M. Adda-Decker and L. Lamel, "Do speech recognizers prefer female speakers?" in *INTERSPEECH*, 2005.

[11] M. Garnerin, S. Rossato, and L. Besacier, "Gender Representation in French Broadcast Corpora and Its Impact on ASR Performance," in *AI4TV@MM 2019*, 2019.

[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. Chizhikova, H. Billinghurst, M. Elizabeth, S. Hossain, A. Kulkarni, G. Guibon, and M. Couceiro, "Factorizing Gender Bias in Automatic Speech Recognition for Mexican Spanish," 2024. [Online]. Available: https://hal.science/hal-04607587

[14] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in *ACL*, 2019.

[15] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau, "Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning," 2020.

[16] S. Budennyy, V. D. Lazarev, N. O. Zakharenko, A. Y. Korovin, O. Plosskaya, D. Dimitrov, V. Arkhipkin, I. Oseledets, I. Barsola, I. M. Egorov, A. Kosterina, and L. Zhukov, "eco2ai: Carbon Emissions Tracking of Machine Learning Models as the First Step Towards Sustainable AI," *Doklady Mathematics*, 2022.

[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*, 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] M. Du, F. Yang, N. Zou, and X. Hu, "Fairness in Deep Learning: A Computational Perspective," *IEEE Intelligent Systems*, 2019.

[21] 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.

[22] R. Tatman and C. Kasten, "Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions," in *INTERSPEECH*, 2017.

[23] A. Kulkarni, A. Tokareva, R. Qureshi, and M. Couceiro, "The Balancing Act: Unmasking and Alleviating ASR Biases in Portuguese," *EACL Workshop LT-EDI*, 2024.

[24] X. Li, F. Metze, D. R. Mortensen, A. W. Black, and S. Watanabe, "ASR2K: Speech Recognition for Around 2000 Languages without Audio," in *INTERSPEECH*, 2022.

[25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *ICML 2023*, ser. Proceedings of Machine Learning Research, 2023.

[26] 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.

[27] 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.

[28] T. Parcollet and M. Ravanelli, "The Energy and Carbon Footprint of Training End-to-End Speech Recognizers," in *INTERSPEECH*, 2021.

[29] I. Lakim, E. Almazrouei, I. Abualhaol, M. Debbah, and J. Launay, "A holistic assessment of the carbon footprint of noor, a very large arabic language model," *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, 2022.

[30] J. Dodge, T. Prewitt, R. T. des Combes, E. Odmark, R. Schwartz, E. Strubell, A. S. Luccioni, N. A. Smith, N. DeCario, and W. Buchanan, "Measuring the carbon intensity of ai in cloud instances," *Proceedings of ACM Conference on Fairness, Accountability, and Transparency*, 2022.

[31] J. Meyer, L. Rauchenstein, J. D. Eisenberg, and N. Howell, "Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in speech applications," in *LREC*, 2020.

[32] B. Porgali, V. Albiero, J. Ryda, C. C. Ferrer, and C. Hazirbas, "The Casual Conversations v2 Dataset," in *CVPR Workshops*, 2023.

[33] 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*, 2021.

[34] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech representations," in *NeurIPS*, 2020.

[35] E. E. Agency, "Greenhouse gas emission intensity of electricity generation in Europe," in *European Environment Agency*, 2020.