Group Fairness in Multilingual Speech Recognition Models

Anna Katrine van Zee
Department of Computer Science
University of Copenhagen
akj@di.ku.dk

Marc van Zee
Google Deepmind
marcvanzee@google.com

Anders Søgaard
Department of Computer Science
Center for Philosophy of AI
University of Copenhagen
soegaard@di.ku.dk

Abstract

We evaluate the performance disparity of the Whisper and MMS families of ASR models across the VoxPopuli and Common Voice multilingual datasets, with an eye toward intersectionality. Our two most important findings are that model size, surprisingly, correlates logarithmically with worst-case performance disparities, meaning that larger (and better) models are less fair. We also observe the importance of intersectionality. In particular, models often exhibit significant performance disparity across binary gender *for adolescents*.

1 Introduction

Automatic speech recognition (ASR) has improved greatly, largely due to representation learning from raw audio. Data scarcity is no longer a major bottleneck for many of the world's languages, and high-quality speech recognition models become more and more integrated in both our private and public lives: From automatically transcribing court proceedings or doctor's notes, to extracting speech from police patrolling, meetings or for hearing aids, speech recognition models have the potential to ease many of the mundane but important tasks we perform on a daily basis.

Performance of ASR models has been shown to vary substantially across user groups (Koenecke et al., 2020; Martin and Tang, 2020; Ngueajio and Washington, 2022). Partial mitigation of performance disparities across user groups is sometimes possible, through distributionally robust optimization (Sagawa* et al., 2020) or spectral decoupling (Pezeshki et al., 2020), for example, but is computationally expensive and requires large amounts of data annotated with demographic information, e.g., protected attributes of speakers. In this study, we show how *small* amounts of such data can be used to evaluate performance disparities, benchmarking two state-of-the-art ASR architectures across languages and demographics.

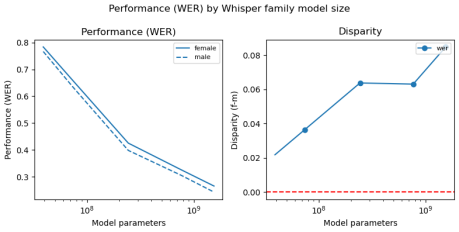

Figure 1: Model performance per binary gender (left) and disparity (right) as a function of model size (log-scale). Dots indicate significant performance disparity ($p<0.05$).

Our purpose is twofold; (i) We want to know what the performance is for protected groups across a variety of speech models; and (ii) we want to create a baseline for other ASR models to compare against. We believe (i) is extremely important, because of the large-scale impact of these models on our everyday lives and the societal imbalances they can reinforce. Establishing a practice around fairness evaluation is important also for future generations of ASR to ensure that benefits are equally distributed across user groups.

Protected Attributes Protected attributes refer to demographic characteristics of individuals such as race, gender, age, and religion, which are considered protected from being used as a basis for discrimination or bias in decision-making processes. In the context of data and machine learning, the consideration of protected attributes becomes crucial to ensure fairness and prevent biases in automated decision-making systems.

If an ASR system is not trained to handle linguistic variation, the system may exhibit much higher error rates for individuals with certain protected attributes –especially if these are correlated with particular accents, dialects, or speech patterns, as is the case of African-American Vernacular English. This can disproportionately impact individuals from specific linguistic or cultural backgrounds,

Findings of the Association for Computational Linguistics: NAACL 2024, pages 2213–2226

leading to unfair outcomes and reduced accessibility for those groups and a perpetuation of existing societal biases and discrimination.

Intersectionality Intersectionality, as coined by Kimberlé Crenshaw (Crenshaw, 1989), illuminates the intricate interconnections among multiple social identities and their role in shaping individuals' experiences and social inequalities. This concept acknowledges that oppression, discrimination, and privilege operate in multidimensional and overlapping ways, defying understanding through singular identity categories. For instance, a woman of color may encounter distinct forms of discrimination differing from those faced by a white woman or a man of color.

The understanding of intersectionality posits that individuals possess a myriad of social identities simultaneously, spanning race, gender, class, sexuality, disability, and more. These identities don't exist in isolation but rather intersect and interact, producing unique challenges and experiences.

Furthermore, intersectionality challenges the simplistic notion that social categories operate independently, highlighting instead the complex interplay between identities. It underscores the interconnected systems of power based on social identities and acknowledges that individuals' experiences of oppression are influenced by the intersections of these identities.

In essence, intersectionality offers a comprehensive framework for comprehending the complexities of identity-based discrimination and privilege, urging for a holistic approach that considers the intertwined systems of power and discrimination. This lens is particularly valuable in examining disparities in AI-driven discrimination or inclusion concerning language use by different social groups.

In the following, we present the data and ASR models we consider for investigating performance disparity between the binary genders and the intersectionality of age and binary gender.

2 Datasets

We make use of two multilingual, open source datasets to evaluate the performance disparity of the two families of ASR models with respect to gender fairness and intersectionality in gender and age.

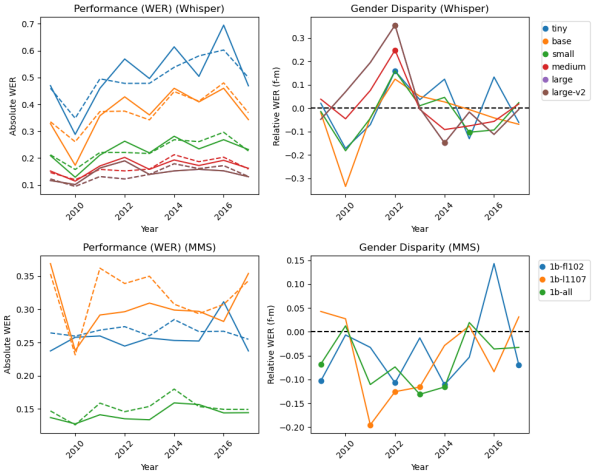

Figure 2: Word error rate (WER) and gender disparity in ASR models for binary genders across years. Left column shows performance results (WER) for Whisper (top) and MMS (bottom) families, and right column is the gap in performance between the binary genders for Whisper (top) and MMS (bottom). Solid lines show performance for female speakers, dashed lines for male. Dots indicate significance ($p \geq 0.01$).

Common Voice[1] is a crowdsourced, continuously developed dataset covering over 200 languages and VoxPopuli[2] is a collection of speeches given in the European Parliament between 2009–2020. Both datasets contain demographic information about the speakers; CommonVoice has gender and age annotations, VoxPopuli has gender markings as well as timestamps for each utterance.

2.1 Limitations and Code

To our knowledge, no available open source dataset exists that would allow us to test the performance disparity for other attributes than binary gender and age or the intersectionality of other attributes than these two. Likewise, we have not been able to find data with nonbinary genders annotated. We redistribute the processed datasets to facilitate hassle-free fairness evaluation for binary gender and age along with open-source evaluation code for testing and visualizing results.[3]

Evaluating Public Models

We evaluate two publicly available ASR model families, namely the Whisper (Radford et al., 2022)

[1] https://huggingface.co/datasets/mozilla-foundation/common_voice_12_0

[2] https://huggingface.co/datasets/facebook/voxpopuli

[3] Anonymized Github link.

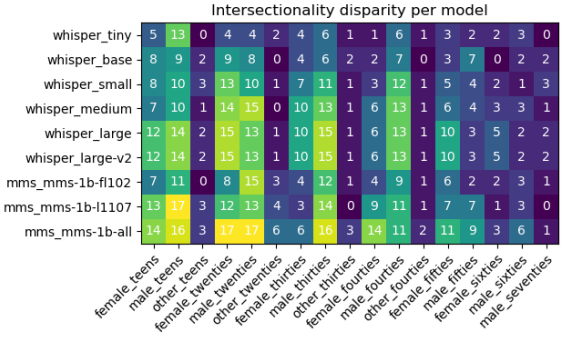

Figure 3: Intersectionality results. We report the number of statistically significant ($p < 0.05$) performance disparities for a particular pair of demographic variables. Performance, again, is measured across multiple languages. We see that, on average, larger models exhibit more intersectionality effects, and we clearly see more disparate performance among younger speakers who identify as men.

and MMS (Pratap et al., 2023) models, i.e., a total of eight models. Both model families consist of multilingual, multitask models. They are also easily accessible models and go-to models for hundreds of companies using ASR in their products.

2.2 Whisper

Whisper is a family of automatic speech recognition (ASR) systems developed by OpenAI. The models are trained on 680,000 hours of web data in 97 languages, and they have parameters ranging from 39M in the 'tiny' model to 1550M in the two 'large' models. Whisper training involves data augmentation, applying transformations to the audio spectrograms during training, including time warping, frequency masking, and time masking. Such data augmentation strategy helps the model generalize better to different acoustic conditions.

2.3 MMS

The Massively Multilingual Speech (MMS) family of ASR models are developed by Meta and trained on 500,000 hours of speech data in 1400+ languages. Based on wav2vec 2.0 models (Baevski et al., 2020), MMS leverages self-supervised methods for learning from a large, new corpus of religious texts. The models all have 1B parameters, but they have been fine-tuned on different datasets.

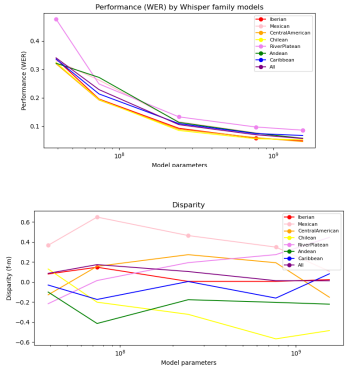

Figure 4: Model performance per accent (top) and performance disparity between genders *within* a dialect (bottom) as a function of model size (log-scale). Dots indicate significant performance disparity (p<0.05).

3 Results

We evaluate all models in the Whisper family (of different sizes) and all models in the MMS family (of different training data) across all demographics in all languages in our two datasets. This is a total of **651** experiments. We then run significance test on all combinations of language, dataset, model, and model size (for the Whisper family). We find significant disparity in performance between the binary genders in 29% of the cases (11% of these negatively for women, 17% for men).

Performance disparity is prevalent across languages and across models, and it seems that model size correlates positively with such disparity (Figure 1). Here, we plot the results with model size on the x-axis, and relative disparity difference on the y-axis. We see that there is a positive, logarithmic correlation between the two variables. Figure 3 shows how gender disparities are particularly high for younger speakers who identify as men. These results showcase how inferring a model's fairness from its parity on data from one demographic group, e.g. adult users, is insufficient.

3.1 A Closer look at Spanish

We zoom in and take a closer look at the performance of the Whisper family models on 7 Spanish dialects.[4] We use the CommonVoice dataset, where gender, age, and dialect are marked for 1829 speakers.

First, we look at the overall performance of the

[4]We exclude the MMS family from this analysis since their performance on Spanish is too poor. The best MMS model (1b-all) is on par with the worst performing Whisper model (tiny).

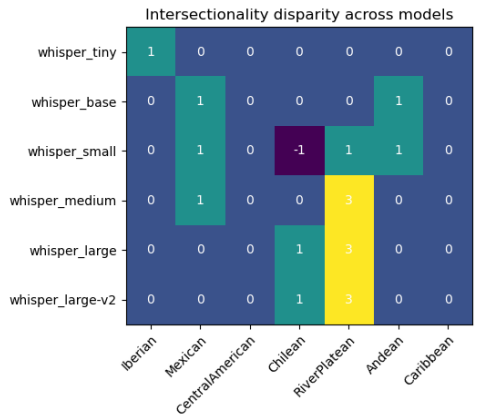

Figure 5: Three-tonged – age, gender, accent – intersectional performance results for Spanish dialects across different Whisper family model sizes. A negative result indicates positive disparate performance.

Whisper model family on the Latin American Spanish dialects and on Iberian Spanish (see Figure 4).[5] We note that performance for all dialects increases (WER decreases) as model size increases, but that the performance is disparate for speakers of River Platean Spanish independent of model size. Performance is not disparate between genders across Spanish.

We then plot the intersectional performance disparity between binary genders *within* each dialect as a function of model size in Figure 4b, ie. female speakers of Mexican Spanish against non-female speakers of Mexican Spanish. We see that while performance increases (lower WER) for all dialects as the model size increases, gender disparity exists in all dialects (except perhaps Iberian), and there is no clear improvement in gender disparity within a dialect when model size increases.

Finally, we investigate the performance disparity across three-tonged intersectional groups with gender, age and accent (see Figure 5). Performance disparity between intersectional groups intensifies with model size, and particularly, female Mexican speakers under 40 and male speakers of Andean under 40 suffer from disparate intersectional performance along with female speakers in their sixties who speak River Platean Spanish. These findings support the two-tonged intersectional results (age and dialect), but indicate that particular age groups are affected by the disparate performance results.

[5]We group the Iberian Spanish dialects together and focus on the Latin American Spanishes in line with the NAACL 2024 theme track.

4 Discussion

Mitigation Some researchers have reported on attempts to make ASR systems less disparate. Boito et al. (2022) report that training ASR models for specific demographic groups did *not* reduce performance disparity. Such strategies also have trouble scaling in light of intersectionality. Veliche and Fung (2023) propose conditioning on cluster IDs with clusters being proxies for demographic groups, but their approach is not easily integrated in pretrained ASR models such as Whisper and MMS. Dheram et al. (2022) had limited success with oversampling from minority groups.

Fairness over Time In ASR research, the predominant focus has been on examining fairness within a static framework, where it is assumed that the data generation process remains constant over time. Nevertheless, these approaches tend to overlook the significant drift in data over time, a phenomenon frequently observed in real-world scenarios. How people talk, and what they talk about, changes over time. What specific demographics talk about changes even faster.

Preliminary investigations have revealed that enforcing static fairness constraints in dynamic systems can lead to inequitable data distributions and, in some cases, exacerbate existing biases (Søgaard et al., 2021). Furthermore, the emergence of powerful large-scale generative models has brought to the forefront the necessity of comprehending fairness within evolving systems. The widespread deployment and versatile capabilities of these models raise a crucial question: how can we assess these models for fairness and effectively mitigate observed biases from a long-term perspective?

As a small step in what we take to be the right direction, we also examined how the performance disparities of Whisper and MMS evolve over time. Since the models are trained on data from the entire period (2009–2017), our protocol does not simulate evaluation on future data, only variance across time. See Figure 2 for an overview. We see a small effect as we depart from the period's average, but with high general variance. The smallest disparities are observed in 2010, 2013, and 2015. Since the VoxPopuli is a collection of speeches from the European Parliament, it is likely that we would see larger variance in datasets from less formal settings. We encourage other researchers to seek out or develop new datasets that can give insights into the variance in performance over time.

Potential Implications The consistent performance gaps observed among demographic groups present a significant challenge for the practical application of ASR models in real-world scenarios. As transcription services for administrative tasks, customer service voicebots, and subtitle creation for recommender systems become ubiquitous across various domains, the disparity in performance across demographics, as demonstrated in this paper, results in certain user groups being underserved. This can consequently lead to users abandoning the service altogether or impose an unjust burden, such as additional manual administration in healthcare, on those belonging to the disadvantaged demographic group. Parate performance, on the other hand, can increase user retention and ameliorate discrimination in the workplace or in access to information.

5 Conclusion

We highlight the potential social impact of ASR's performance disparities across demographic groups in the –to our knowledge– first study of its kind. We run a total of **651** experiments evaluating state-of-the-art model families on data containing protected attributes, namely binary gender and age. We release the curated dataset to ease implementation of disparity testing for researchers and developers. Our main findings were as follows: (i) Larger models surprisingly exhibit more performance disparity. (ii) Intersectional effects are evident, largely affecting the younger speakers who identify as men. (iii) Finally, we see small signs of temporal variation in disparity figures, but less dramatic than the variation observed across protected attributes.

Future Directions Our examination of performance disparities among demographic groups in ASR systems represents an initial exploration of a technology increasingly relied upon across various sectors and applications worldwide. We anticipate that numerous similar investigations will ensue, as numerous questions regarding differential performance among groups remain unanswered. While awaiting longitudinal data to fully grasp the implications of ASR performance on discrimination and racism beyond the system itself, we urge researchers and developers to prioritize examining performance for children. This demographic, often underrepresented in research yet overrepresented in platforms like social media, is unique in that they are learning language use while engaging with ASR systems, unlike previous generations. Ensuring optimal performance for this demographic is of utmost importance.

Acknowledgements

This work was supported by Innovationsfonden (grant number 2081-00022A). We thank Jasmijn Bastings for her useful comments.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Marcely Zanon Boito, Laurent Besacier, Natalia A. Tomashenko, and Yannick Estève. 2022. A study of gender impact in self-supervised models for speech-to-text systems. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 1278–1282. ISCA.

Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum*, 140:139–167.

Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke. 2022. Toward fairness in speech recognition: Discovery and mitigation of performance disparities. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 1268–1272. ISCA.

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. *PNAS*, 117(14):7684–7689.

Joshua Martin and Kevin Tang. 2020. Understanding racial disparities in automatic speech recognition: The case of habitual "be".

Mikel K. Ngueajio and Gloria Washington. 2022. Hey asr system! why aren't you more inclusive? automatic speech recognition systems' bias and proposed bias mitigation techniques. a literature review. In *HCI*.

Mohammad Pezeshki, Sekouba Kaba, Yoshua Bengio, Aaron C. Courville, Doina Precup, and Guillaume Lajoie. 2020. Gradient starvation: A learning proclivity in neural networks. In *Neural Information Processing Systems*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *arXiv*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *International Conference on Learning Representations*.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

Irina-Elena Veliche and Pascale Fung. 2023. Improving fairness and robustness in end-to-end speech recognition through unsupervised clustering. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Appendix A Intersectionality Results

Below, we provide the female and male performance results (WER) for all Whisper and MMS models on the CommonVoice and VoxPopuli datasets for all languages. For intersectionality results for all models on the CommonVoice dataset, please see the project's Github page.

Appendix B Female/Male Performance (Word Error Rate) for Whisper models on CommonVoice

Language	female	male	diff
az		1.149	
es	0.375	0.379	-0.004
nl	0.484	0.525	-0.041
da	0.772	0.822	-0.05
ro	0.817	0.843	-0.026
sw	1.383	1.246	0.137
hy-AM	1.566	1.945	-0.379
fi	1.247	0.959	0.288
ba	1.479	1.615	-0.136
cs	0.868	0.895	-0.027
it	0.526	0.527	-0.001
pl	0.572	0.586	-0.014
cy	1.304	1.645	-0.341
el	0.899	0.752	0.147
bg	0.843	0.85	-0.007
th	1.351	1.306	0.045
zh-HK	2.066	1.303	0.763
uz	1.631	1.732	-0.101
ha	0.953	1.087	-0.134
sv-SE	0.64	0.669	-0.029
ca	0.631	0.662	-0.031
lv	0.887	0.902	-0.015
eu	1.012	1.097	-0.085
et	1.025	1.06	-0.035
br	1.045	1.162	-0.117
pt	0.532	0.487	0.045
hu	1.071	1.044	0.027
zh-TW	0.87	0.904	-0.034
mn	2.324	2.443	-0.119
kk		3.944	
fa	1.817	2.089	-0.272
en	0.312	0.339	-0.027
mt	1.09	1.178	-0.088
ka	2.316	2.252	0.064
sk	1.256	1.252	0.004
zh-CN	1.202	1.434	-0.232
ar	0.99	1.042	-0.052
as		1.246	
mk		0.754	
tr	0.77	0.693	0.077
uk	0.715	0.656	0.059
gl	0.615	0.642	-0.027
pa-IN	1	1.483	-0.483
nn-NO		1.188	
sr	1.052	1.208	-0.156
fr	0.638	0.598	0.04
ml		1.414	
id	0.607	0.612	-0.005
vi	0.7	0.933	-0.233
tt	1.791	1.317	0.474
ja	1.035	1.322	-0.287
lt	1.077	1.069	0.008
de	0.52	0.436	0.084
ta	1.247	1.423	-0.176
bn	1.565	1.363	0.202
ru	0.444	0.513	-0.069
ur	1.123	1.337	-0.214
mr	1.117	1.095	0.022
be	1.058	1.022	0.036
sl	0.856	0.914	-0.058
hi	1.224	1.291	-0.067

Table 1: Word error rate (WER) with Whisper Tiny on CommonVoice

Language	female	male	diff
be	0.945	0.956	-0.011
mr	1.128	1.037	0.091
hi	1.077	1.324	-0.247
sl	0.761	0.746	0.015
ru	0.284	0.341	-0.057
bn	1.224	1.232	-0.008
ta	0.524	0.615	-0.091
ur	0.847	0.794	0.053
lt	0.996	0.96	0.036
ja	0.981	1.089	-0.108
de	0.358	0.317	0.041
sr	1.024	1.127	-0.103
id	0.504	0.451	0.053
vi	0.657	0.521	0.136
tt	1.352	1.446	-0.094
ml		1.038	
fr	0.507	0.428	0.079
gl	0.56	0.519	0.041
uk	0.578	0.542	0.036
ar	0.985	0.948	0.037
as		1.256	
tr	0.512	0.48	0.032
mk		0.694	
ka	1.63	1.687	-0.057
pa-IN	1	1.131	-0.131
mt	1.544	1.651	-0.107
nn-NO		0.723	
sk	1.444	1.005	0.439
fa	1.524	1.295	0.229
zh-CN	1.249	1.202	0.047
kk		1.696	
en	0.234	0.231	0.003
hu	1.123	0.95	0.173
sv-SE	0.497	0.52	-0.023
mn	3.634	3.55	0.084
eu	0.966	0.979	-0.013
et	0.964	1.027	-0.063
lv	0.89	0.835	0.055
ca	0.452	0.627	-0.175
pt	0.393	0.407	-0.014
zh-HK	1.526	1.564	-0.038
br	1.207	1.438	-0.231
zh-TW	0.731	0.837	-0.106
ha	3.031	2.819	0.212
th	0.783	0.783	0
bg	0.858	0.842	0.016
uz	3.259	2.787	0.472
pl	0.446	0.429	0.017
it	0.396	0.385	0.011
hy-AM	1.968	2.596	-0.628
el	0.636	0.651	-0.015
cy	1.196	1.341	-0.145
ba	1.711	1.767	-0.056
fi	0.765	0.602	0.163
cs	0.73	0.723	0.007
sw	1.604	1.417	0.187
nl	0.342	0.369	-0.027
es	0.286	0.272	0.014
az		0.707	
ro	0.689	0.709	-0.02
da	0.594	0.643	-0.049

Table 2: Word error rate (WER) with Whisper Base on CommonVoice

Language	female	male	diff
ta	0.287	0.346	-0.059
bn	1.282	1.395	-0.113
ru	0.145	0.196	-0.051
ur	0.398	0.487	-0.089
hy-AM	1.187	2.031	-0.844
mr	0.641	0.714	-0.073
be	0.868	0.859	0.009
sl	0.488	0.495	-0.007
hi	0.748	0.586	0.162
sr	1.048	1.017	0.031
fr	0.305	0.271	0.034
ml		1.249	
vi	0.442	0.304	0.138
id	0.21	0.228	-0.018
tt	1.077	1.095	-0.018
ja	0.835	0.903	-0.068
lt	0.881	0.797	0.084
de	0.218	0.166	0.052
ar	0.593	0.56	0.033
as		1.488	
mk		0.484	
sv-SE	0.264	0.268	-0.004
tr	0.293	0.276	0.017
uk	0.37	0.323	0.047
gl	0.371	0.373	-0.002
zh-HK	1.024	1.121	-0.097
kk		1.34	
fa	0.754	0.879	-0.125
zh-TW	0.535	0.578	-0.043
en	0.178	0.165	0.013
mt	1.007	1.032	-0.025
ka	5.378	5.952	-0.574
sk	0.843	0.76	0.083
ca	0.258	0.38	-0.122
lv	0.629	0.643	-0.014
eu	0.887	0.873	0.014
et	0.714	0.745	-0.031
br	3.116	1.988	1.128
pt	0.279	0.192	0.087
hu	0.561	0.578	-0.017
mn	1.64	1.966	-0.326
bg	0.454	0.534	-0.08
pa-IN	1	1.242	-0.242
th	0.584	0.522	0.062
uz	1.297	1.16	0.137
nn-NO		1.33	
zh-CN	1.101	1.046	0.055
ha	0.875	0.894	-0.019
fi	0.514	0.319	0.195
ba	1.251	1.262	-0.011
cs	0.394	0.403	-0.009
it	0.175	0.202	-0.027
pl	0.249	0.244	0.005
cy	0.767	0.775	-0.008
el	0.395	0.389	0.006
az		0.615	
es	0.142	0.121	0.021
nl	0.179	0.183	-0.004
da	0.369	0.402	-0.033
ro	0.378	0.416	-0.038
sw	0.991	0.96	0.031

Table 3: Word error rate (WER) with Whisper Small on CommonVoice

Language	female	male	diff
ur	0.311	0.361	-0.05
ta	0.198	0.252	-0.054
bn	1.161	1.261	-0.1
ru	0.089	0.107	-0.018
sl	0.297	0.321	-0.024
hi	0.298	0.34	-0.042
mr	0.533	0.59	-0.057
be	0.678	0.684	-0.006
fr	0.196	0.181	0.015
ml		1	
tt	1.135	1.047	0.088
id	0.113	0.132	-0.019
vi	0.357	0.203	0.154
sr	0.891	0.812	0.079
hy-AM	0.572	0.786	-0.214
de	0.137	0.104	0.033
ja	0.756	0.803	-0.047
lt	0.675	0.535	0.14
mk		0.333	
tr	0.261	0.191	0.07
ar	0.495	0.469	0.026
as		1.047	
zh-TW	0.481	0.455	0.026
uk	0.243	0.211	0.032
gl	0.253	0.238	0.015
zh-HK	0.88	0.952	-0.072
en	0.138	0.138	0
kk		0.589	
fa	0.592	0.744	-0.152
sk	0.566	0.547	0.019
sv-SE	0.185	0.173	0.012
mt	0.942	0.917	0.025
ka	1.22	1.322	-0.102
br	1.308	1.13	0.178
pt	0.192	0.127	0.065
zh-CN	0.935	0.912	0.023
ca	0.175	0.25	-0.075
lv	0.449	0.437	0.012
eu	0.671	0.657	0.014
et	0.486	0.521	-0.035
mn	1.489	1.454	0.035
nn-NO		0.368	
hu	0.331	0.371	-0.04
pa-IN	1	1.135	-0.135
uz	1.634	1.636	-0.002
bg	0.253	0.33	-0.077
th	0.347	0.506	-0.159
ha	1.131	1.515	-0.384
cs	0.223	0.233	-0.01
fi	0.354	0.175	0.179
ba	1.281	1.255	0.026
cy	0.481	0.514	-0.033
el	0.206	0.242	-0.036
it	0.099	0.114	-0.015
pl	0.14	0.134	0.006
da	0.219	0.266	-0.047
ro	0.219	0.264	-0.045
az		0.444	
es	0.088	0.092	-0.004
nl	0.088	0.103	-0.015
sw	0.743	0.682	0.061

Table 4: Word error rate (WER) with Whisper Medium on CommonVoice

Language	female	male	diff
fi	0.289	0.147	0.142
sv-SE	0.131	0.138	-0.007
ba	1.112	1.127	-0.015
cs	0.156	0.173	-0.017
zh-HK	0.876	0.908	-0.032
it	0.088	0.092	-0.004
pl	0.105	0.107	-0.002
cy	0.363	0.383	-0.02
el	0.173	0.201	-0.028
az		0.329	
zh-TW	0.583	0.431	0.152
es	0.071	0.069	0.002
nl	0.065	0.073	-0.008
da	0.164	0.214	-0.05
ro	0.148	0.174	-0.026
sw	0.613	0.582	0.031
ca	0.156	0.187	-0.031
lv	0.33	0.316	0.014
hy-AM	0.483	0.694	-0.211
eu	0.509	0.525	-0.016
et	0.344	0.396	-0.052
br	1.087	1.139	-0.052
pt	0.164	0.102	0.062
hu	0.288	0.274	0.014
mn	1.357	1.357	0
bg	0.179	0.26	-0.081
th	0.259	0.324	-0.065
uz	1.008	0.968	0.04
ha	0.896	0.982	-0.086
ar	0.423	0.385	0.038
as		1.053	
mk		0.228	
tr	0.186	0.156	0.03
uk	0.181	0.162	0.019
gl	0.204	0.187	0.017
kk		0.63	
fa	0.443	0.473	-0.03
en	0.121	0.116	0.005
mt	1.009	0.879	0.13
ka	1.236	1.206	0.03
sk	0.532	0.424	0.108
ta	0.173	0.204	-0.031
bn	1.061	1.045	0.016
ru	0.068	0.085	-0.017
ur	0.272	0.331	-0.059
mr	0.368	0.374	-0.006
be	0.521	0.536	-0.015
sl	0.22	0.253	-0.033
hi	0.166	0.252	-0.086
sr	0.711	0.726	-0.015
nn-NO		0.434	
fr	0.174	0.153	0.021
ml		1.389	
pa-IN	1	1.031	-0.031
tt	1.428	1.2	0.228
vi	0.301	0.192	0.109
id	0.087	0.097	-0.01
ja	0.688	0.754	-0.066
lt	0.486	0.386	0.1
zh-CN	0.973	0.918	0.055
de	0.105	0.077	0.028

Table 5: Word error rate (WER) with Whisper Large on CommonVoice

Appendix C Female/Male Performance (Word Error Rate) for MMS models on CommonVoice

Language	female	male	diff
nn-NO		0.434	
pa-IN	1	1.031	-0.031
ha	0.896	0.982	-0.086
bg	0.179	0.26	-0.081
th	0.259	0.324	-0.065
uz	1.008	0.968	0.04
zh-CN	0.973	0.918	0.055
hu	0.288	0.274	0.014
mn	1.357	1.357	0
lv	0.33	0.316	0.014
et	0.344	0.396	-0.052
eu	0.509	0.525	-0.016
ca	0.156	0.187	-0.031
pt	0.164	0.102	0.062
br	1.087	1.139	-0.052
sw	0.613	0.582	0.031
es	0.071	0.069	0.002
nl	0.065	0.073	-0.008
az		0.329	
da	0.164	0.214	-0.05
ro	0.148	0.174	-0.026
pl	0.105	0.107	-0.002
it	0.088	0.092	-0.004
el	0.173	0.201	-0.028
cy	0.363	0.383	-0.02
ba	1.112	1.127	-0.015
fi	0.289	0.147	0.142
cs	0.156	0.173	-0.017
lt	0.486	0.386	0.1
ja	0.688	0.754	-0.066
de	0.105	0.077	0.028
sr	0.711	0.726	-0.015
tt	1.428	1.2	0.228
id	0.087	0.097	-0.01
vi	0.301	0.192	0.109
fr	0.174	0.153	0.021
ml		1.389	
be	0.521	0.536	-0.015
hy-AM	0.483	0.694	-0.211
mr	0.368	0.374	-0.006
hi	0.166	0.252	-0.086
sl	0.22	0.253	-0.033
ru	0.068	0.085	-0.017
ta	0.173	0.204	-0.031
bn	1.061	1.045	0.016
ur	0.272	0.331	-0.059
ka	1.236	1.206	0.03
zh-TW	0.583	0.431	0.152
mt	1.009	0.879	0.13
sk	0.532	0.424	0.108
fa	0.443	0.473	-0.03
kk		0.63	
en	0.121	0.116	0.005
gl	0.204	0.187	0.017
sv-SE	0.131	0.138	-0.007
uk	0.181	0.162	0.019
as		1.053	
ar	0.423	0.385	0.038
zh-HK	0.876	0.908	-0.032
tr	0.186	0.156	0.03
mk		0.228	

Table 6: Word error rate (WER) with Whisper Large-v2 on CommonVoice

Language	female	male	diff
bn	0.414	0.511	-0.097
ta	0.594	0.685	-0.091
ru	0.599	0.582	0.017
mr	0.488	0.525	-0.037
be	0.464	0.476	-0.012
sl	0.619	0.621	-0.002
hi	0.415	0.441	-0.026
ml		0.636	
fr	0.581	0.58	0.001
ky	0.556	0.593	-0.037
id	0.568	0.574	-0.006
vi	0.772	0.538	0.234
ja	1.999	2.22	-0.221
lt	0.559	0.535	0.024
de	0.625	0.576	0.049
as		0.569	
ar	1.075	1.076	-0.001
mk		0.3	
tr	0.746	0.71	0.036
uk	0.618	0.62	-0.002
gl	0.488	0.43	0.058
kk		0.754	
fa	1.064	1.057	0.007
en	0.59	0.649	-0.059
mt	0.52	0.506	0.014
ka	0.39	0.45	-0.06
sk	1.066	0.98	0.086
lg	0.663	0.649	0.014
ca	0.483	0.481	0.002
et	0.449	0.495	-0.046
lv	0.711	0.671	0.04
pt	0.858	0.627	0.231
hu	0.564	0.634	-0.07
mn	0.652	0.589	0.063
ig		0.714	
th	0.841	0.885	-0.044
bg	1.022	1.013	0.009
ha	0.541	0.522	0.019
fi	0.636	0.533	0.103
cs	0.691	0.7	-0.009
it	0.422	0.438	-0.016
pl	0.535	0.534	0.001
cy	0.606	0.647	-0.041
el	1.063	1.113	-0.05
nl	0.362	0.377	-0.015
es	0.45	0.482	-0.032
ro	0.412	0.413	-0.001
da	0.454	0.446	0.008

Table 7: Word error rate (WER) with MMS-MMS-1b-fl102 on CommonVoice

Language	female	male	diff
ta	0.606	0.662	-0.056
bn	0.531	0.606	-0.075
ru	0.603	0.621	-0.018
cv	0.712	0.776	-0.064
hi	0.445	0.483	-0.038
mr	0.547	0.587	-0.04
fr	0.566	0.551	0.015
ml		0.614	
tt	0.705	0.72	-0.015
ky	0.605	0.635	-0.03
id	0.528	0.534	-0.006
vi	0.683	0.5	0.183
de	0.635	0.594	0.041
tr	0.73	0.708	0.022
ar	0.57	0.54	0.03
as		0.614	
uk	0.634	0.629	0.005
en	0.549	0.582	-0.033
kk		0.727	
fa	0.548	0.581	-0.033
lg	0.566	0.55	0.016
dv	0.657	0.644	0.013
pt	0.569	0.524	0.045
ca	0.488	0.491	-0.003
lv	0.664	0.649	0.015
eu	0.531	0.532	-0.001
mn	0.744	0.668	0.076
hu	0.591	0.688	-0.097
bg	0.484	0.512	-0.028
th	0.938	0.938	0
ha	0.55	0.524	0.026
rw	0.543	0.609	-0.066
fi	0.66	0.564	0.096
ba	0.622	0.661	-0.039
cy	0.667	0.649	0.018
el	0.539	0.616	-0.077
pl	0.54	0.536	0.004
ro	0.395	0.411	-0.016
gn	0.827	0.867	-0.04
es	0.397	0.41	-0.013
nl	0.401	0.418	-0.017

Table 8: Word error rate (WER) with MMS-MMS-1b-l1107 on CommonVoice

Language	female	male	diff
th	0.802	0.843	-0.041
bg	0.455	0.474	-0.019
ab	0.77	0.765	0.005
ha	0.557	0.488	0.069
rw	0.482	0.544	-0.062
br	0.774	0.808	-0.034
pt	0.489	0.455	0.034
ca	0.434	0.432	0.002
eu	0.495	0.495	0
et	0.421	0.465	-0.044
lv	0.652	0.634	0.018
mn	0.619	0.547	0.072
ig		0.571	
hu	0.51	0.573	-0.063
ro	0.378	0.382	-0.004
da	0.405	0.42	-0.015
nl	0.362	0.342	0.02
es	0.361	0.376	-0.015
gn	0.72	0.805	-0.085
ia	0.502	0.453	0.049
cs	0.449	0.46	-0.011
fi	0.611	0.509	0.102
ba	0.584	0.615	-0.031
cy	0.561	0.573	-0.012
el	0.461	0.485	-0.024
it	0.388	0.4	-0.012
pl	0.52	0.482	0.038
ml		0.706	
fr	0.436	0.437	-0.001
vi	0.709	0.438	0.271
id	0.512	0.511	0.001
ky	0.516	0.544	-0.028
tt	0.62	0.637	-0.017
de	0.557	0.535	0.022
ja	0.997	0.997	0
lt	0.529	0.513	0.016
bn	0.425	0.505	-0.08
ta	0.533	0.604	-0.071
ru	0.505	0.505	0
sl	0.499	0.51	-0.011
hi	0.362	0.395	-0.033
cv	0.644	0.72	-0.076
mr	0.46	0.493	-0.033
be	0.421	0.433	-0.012
eo	0.439	0.429	0.01
en	0.455	0.478	-0.023
kk		0.696	
fa	0.422	0.444	-0.022
lg	0.507	0.495	0.012
sk	0.532	0.471	0.061
dv	0.495	0.486	0.009
mt	0.493	0.478	0.015
ka	0.359	0.417	-0.058
mk		0.29	
tr	0.698	0.647	0.051
ar	0.547	0.52	0.027
as		0.534	
uk	0.571	0.562	0.009
gl	0.382	0.356	0.026

Table 9: Word error rate (WER) with MMS-MMS-1b-all on CommonVoice

Appendix D Female/Male Performance (Word Error Rate) for Whisper models on VoxPopuli

Language	female	male	diff
fr	0.351	0.359	-0.008
de	0.302	0.409	-0.107
lt	1.121	1.187	-0.066
sl	0.868	0.878	-0.01
en	0.134	0.117	0.017
sk	0.864	0.863	0.001
et	1.02	1.138	-0.118
hu	0.921	1.052	-0.131
ro	0.769	0.793	-0.024
nl	0.471	0.537	-0.066
es	0.343	0.27	0.073
hr	0.796	0.892	-0.096
cs	0.822	0.813	0.009
fi	0.81	0.885	-0.075
it	0.449	0.556	-0.107
pl	0.498	0.47	0.028

Table 10: Word error rate (WER) with Whisper Tiny on VoxPopuli

Language	female	male	diff
it	0.299	0.441	-0.142
pl	0.32	0.305	0.015
cs	0.593	0.782	-0.189
fi	0.448	0.513	-0.065
hr	0.655	0.724	-0.069
ro	0.569	0.602	-0.033
es	0.168	0.178	-0.01
nl	0.351	0.532	-0.181
hu	0.883	0.773	0.11
et	1.061	0.822	0.239
sk	0.68	0.754	-0.074
en	0.116	0.093	0.023
sl	0.679	0.856	-0.177
de	0.206	0.285	-0.079
lt	0.882	1.311	-0.429
fr	0.272	0.347	-0.075

Table 11: Word error rate (WER) with Whisper Base on VoxPopuli

Language	female	male	diff
sl	0.45	0.688	-0.238
fr	0.145	0.158	-0.013
lt	0.604	0.637	-0.033
de	0.144	0.193	-0.049
en	0.105	0.081	0.024
sk	0.395	0.39	0.005
et	1.09	0.621	0.469
hu	0.414	0.431	-0.017
fi	0.317	0.305	0.012
cs	0.306	0.338	-0.032
pl	0.232	0.19	0.042
it	0.229	0.395	-0.166
nl	0.249	0.353	-0.104
es	0.133	0.121	0.012
ro	0.359	0.345	0.014
hr	0.365	0.445	-0.08

Table 12: Word error rate (WER) with Whisper Small on VoxPopuli

Language	female	male	diff
fi	0.184	0.179	0.005
cs	0.202	0.252	-0.05
it	0.18	0.299	-0.119
pl	0.122	0.117	0.005
nl	0.177	0.188	-0.011
es	0.111	0.091	0.02
ro	0.228	0.228	0
hr	0.289	0.311	-0.022
et	0.445	0.411	0.034
hu	0.283	0.255	0.028
en	0.1	0.076	0.024
sk	0.235	0.238	-0.003
sl	0.323	0.626	-0.303
fr	0.115	0.117	-0.002
lt	0.316	0.447	-0.131
de	0.105	0.155	-0.05

Table 13: Word error rate (WER) with Whisper Medium on VoxPopuli

Language	female	male	diff
et	0.307	0.304	0.003
hu	0.261	0.196	0.065
ro	0.169	0.169	0
nl	0.15	0.16	-0.01
es	0.093	0.079	0.014
hr	0.219	0.257	-0.038
cs	0.114	0.153	-0.039
fi	0.167	0.153	0.014
it	0.171	0.244	-0.073
pl	0.112	0.091	0.021
fr	0.111	0.11	0.001
de	0.1	0.151	-0.051
lt	0.23	0.405	-0.175
sl	0.221	0.377	-0.156
en	0.098	0.072	0.026
sk	0.154	0.166	-0.012

Table 14: Word error rate (WER) with Whisper Large on VoxPopuli

Language	female	male	diff
it	0.171	0.244	-0.073
pl	0.112	0.091	0.021
cs	0.114	0.153	-0.039
fi	0.167	0.153	0.014
hr	0.219	0.257	-0.038
ro	0.169	0.169	0
nl	0.15	0.16	-0.01
es	0.093	0.079	0.014
hu	0.261	0.196	0.065
et	0.307	0.304	0.003
sk	0.154	0.166	-0.012
en	0.098	0.072	0.026
sl	0.221	0.377	-0.156
de	0.1	0.151	-0.051
lt	0.23	0.405	-0.175
fr	0.111	0.11	0.001

Table 15: Word error rate (WER) with Whisper Large-v2 on VoxPopuli

Appendix E Performance (Word Error Rate) for MMS models on VoxPopuli

Language	female	male	diff
et	0.298	0.286	0.012
hu	0.33	0.294	0.036
nl	0.316	0.375	-0.059
es	0.311	0.312	-0.001
ro	0.228	0.229	-0.001
hr	0.235	0.294	-0.059
fi	0.282	0.248	0.034
cs	0.186	0.216	-0.03
it	0.242	0.327	-0.085
pl	0.292	0.274	0.018
fr	0.339	0.371	-0.032
lt	0.528	0.603	-0.075
de	0.274	0.371	-0.097
sl	0.232	0.391	-0.159
en	0.356	0.355	0.001
sk	0.465	0.464	0.001

Table 16: Word error rate (WER) with mms-mms-1b-fl102 on VoxPopuli

Language	female	male	diff
en	0.322	0.309	0.013
de	0.788	0.808	-0.02
fr	0.312	0.328	-0.016
pl	0.296	0.295	0.001
fi	0.34	0.313	0.027
es	0.185	0.17	0.015
nl	0.353	0.381	-0.028
ro	0.29	0.298	-0.008
hu	0.323	0.305	0.018

Table 17: Word error rate (WER) with mms-mms-1b-l1107 on VoxPopuli

Language	female	male	diff
cs	0.138	0.167	-0.029
fi	0.223	0.186	0.037
pl	0.168	0.163	0.005
it	0.197	0.262	-0.065
ro	0.131	0.143	-0.012
es	0.137	0.128	0.009
nl	0.212	0.223	-0.011
hr	0.135	0.178	-0.043
et	0.227	0.221	0.006
hu	0.2	0.165	0.035
en	0.148	0.152	-0.004
sk	0.12	0.14	-0.02
sl	0.181	0.303	-0.122
fr	0.157	0.169	-0.012
de	0.165	0.218	-0.053
lt	0.491	0.542	-0.051

Table 18: Word error rate (WER) with mms-mms-1b-all on VoxPopuli