Group Distributionally Robust Optimization for Automatic Speech Recognition

Jacob Mejia
Stanford University
jamejia@stanford.edu

Eric Tang
Stanford University
eatang@stanford.edu

Abstract

In this paper, we improve the robustness of automatic speech recognition models on subsets of the ML-SUPERB benchmark by applying Group Distributionally Robust Optimization (Group-DRO). The ML-SUPERB (Shi et al., 2023) benchmark consists of a variety of multilingual speech recognition datasets, including Multilingual Librispeech, Commonvoice, Fleurs, and more (Ardila et al., 2019; Pratap et al., 2020; Conneau et al., 2023). The diversity of the benchmark, both in terms of low-resource language data as well as in dataset source and collection methods, makes it an interesting challenge for studying generalization and robustness in automatic speech recognition models. Via experiments on the Bantu family of languages, and on English audio across various ML-SUPERB datasets, we show that Group-DRO can help improve worst case performance across more challenging groups while maintaining high average performance.

1 Introduction

Automatic Speech Recognition (ASR) has emerged as an impactful technology showing up in applications such as virtual-assistants, transcription services, customer service and more. ASR models like Wav2Vec2 (Baevski et al., 2020), XLSR (Conneau et al., 2020), and Whisper (Radford et al., 2023) have made strong progress in transcribing spoken language, particularly English, into written text with high accuracy, making ASR an integral area of focus for high impact real-world applications. However, despite the success of such models on high-resource languages (languages that have substantial linguistic resources and support) such as English, ASR models face many challenges with regards to robustness across diverse languages, cultures, and accents.

As a result, a key issue in ASR's deployment is their performance across these unique, diverse groups. Since commonly deployed ASR backbones such as Whisper are typically trained using standard empirical risk minimization techniques, in which loss metrics are averaged across mini-batches during training time (Radford et al., 2023), the performance of these models on low-resource languages and data sources with fewer training examples is naturally downweighted. This can lead to inequitable user experiences for users belonging to these groups, which is more and more undesirable the more ASR technology advances and gradually begins to replace human in the loop systems.

To circumvent these inequities, we can view low-resource languages and uncommon data sources as out of distribution training examples that our ASR models need to be able to generalize to. By overfitting to higher resource settings, ASR models learn spurious correlations that may not hold for atypical groups. Group Distributionally Robust Optimization (Group-DRO) (Sagawa et al., 2019) is an optimization technique proposed to help resolve this disparity in atypical group performance by optimizing over a weighted worst case group loss rather than an average empirical risk minimization loss. Group-DRO has been shown to lead to more robust models over image recognition and text classification datasets like CelebA (Liu et al., 2018), Waterbirds (Sagawa et al., 2019), and MultiNLI (Williams et al., 2017).

In this paper, we apply Group-DRO to speech recognition, an unexplored application in the original Group-DRO paper. Our work aims to improve the robustness of ASR systems, especially for low-resource groups that traditionally experience low test-time performance. Our experiments and conclusions demonstrate Group-DRO's effectiveness in building more reliant ASR systems and we hope that others will continue to explore its application in this sub-field.

2 Related Works

2.1 Distribution Shift and Robust Optimization

The primary source that we draw inspiration from for our work is Group-DRO (Sagawa et al., 2019), which iterates on prior work in methods for robust optimization (Duchi et al., 2021; Ben-Tal et al., 2013), and shows that optimizing over the worst case group loss, combined with stronger regularization techniques such as a heavier ℓ_2 loss or early stopping, can provide robustness benefits for over-parameterized neural network models. Group-DRO is proposed as an alternative to the standard machine learning paradigm of empirical risk minimization (ERM), in which average loss over all samples is minimized.

$$\hat{\theta}_{\text{ERM}} := \arg\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\hat{P}}[\ell(\theta;(x,y))], \tag{1}$$

Group-DRO is defined as follows:

$$\hat{\theta}_{\text{DRO}} := \arg\min_{\theta\in\Theta}\Big\{\hat{\mathcal{R}}(\theta) := \max_{g\in\mathcal{G}} \mathbb{E}_{(x,y)\sim\hat{P}_g}[\ell(\theta;(x,y))]\Big\}, \tag{2}$$

Another potential alternative to ERM that has been explored is Invariant Risk Minimization, which is a learning algorithm proposed by (Arjovsky et al., 2019), in which the goal is to learn representations such that the optimal classifier on top of that representation matches for all environments.

2.2 Datasets

ML-SUPERB is a large scale multilingual benchmark consisting of existing datasets and evaluated on state-of-the-art models including XLSR, Wav2Vec2, Whisper, and more (Conneau et al., 2020; Baevski et al., 2020; Radford et al., 2023). ML-SUPERB aims to provide a robust and standardized benchmark for evaluating speech understanding models, and addresses the need for a comprehensive evaluation framework that encompasses a wide range of speech processing tasks across different languages. It is used to test the performance of a multitude of speech processing tasks including: automatic speech recognition, speaker identification, and emotion recognition. Results across different models/tasks suggest where improvements can be made and offer more insight into how to better develop speech processing systems.

3 Data Exploration

For our data, we use the ML-SUPERB dataset, consisting of over 143 languages across multiple sources. The dataset is structured as follows:

```
- [Data_Source_Name1]
   - [Lang_ID1]
        - transcript_10min_dev.txt
        - transcript_10min_test.txt
        - transcript_10min_train.txt
        - transcript_1h_train.txt
        - wav
        - [Data_Source_Name1]_[Lang_ID1]_000001.wav
        - [Data_Source_Name1]_[Lang_ID1]_000002.wav
            - ...
   - [Lang_ID2]
        - transcript_10min_dev.txt
        - transcript_10min_test.txt
        - transcript_10min_train.txt
        - transcript_1h_train.txt
        - wav
        - [Data_Source_Name1]_[Lang_ID2]_000001.wav
        - [Data_Source_Name1]_[Lang_ID2]_000002.wav
            - ...
   - ...
```

In Figure 1, we show the drastic difference in the number of training samples between high and low-resource languages, with ML-SUPERB consisting of more than 6000 English utterances, but only around 150 training utterances for low-resource languages like Umbundu and Lingala. In Figure 2, we show the datasets that makeup the parts of ML-SUPERB. We can see that almost half of the examples in the dataset come from the Common-voice dataset (Ardila et al., 2019), which could potentially lead to models being evaluated on ML-SUPERB overfitting to the specific distribution of audio samples provided in that dataset. We plan to further investigate dataset source as a type of distribution shift, and whether or not we can potentially apply Group-DRO for improving performance on datasets with fewer examples while maintaining high average performance.

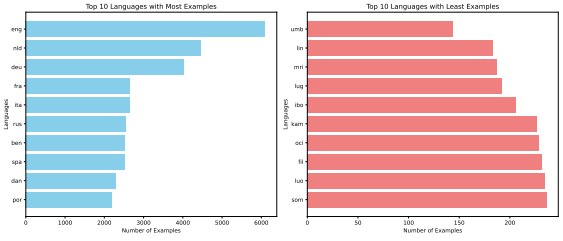

Figure 1: Top and bottom 10 languages by number of training examples.

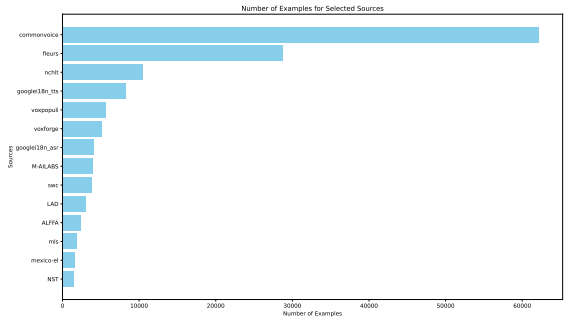

Figure 2: ML-SUPERB datasets sorted by number of training examples.

4 Methods

4.1 Models

We choose to use the pretrained Whisper (Radford et al., 2023) family of models from OpenAI for our experiments, specifically the tiny, base, and small Whisper models. We choose smaller models due to compute resource limitations. These are pretrained on 680k hours of labeled speech data, with 117k hours coming from varying low and high-resource languages and the rest coming from English. We use these models for both zero-shot learning as well as finetuning on the ERM + Group-DRO optimization objective.

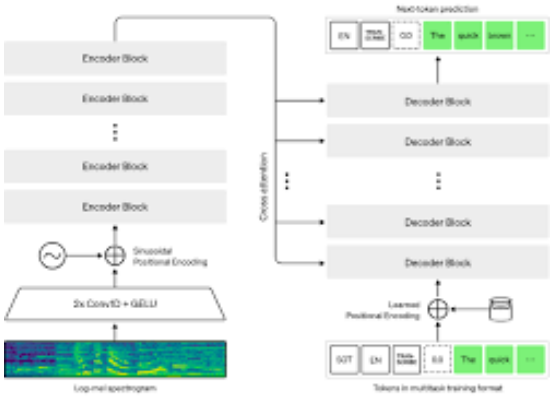

Figure 3: Whisper model architecture from OpenAI.

4.2 Group-DRO Loss

Our approach focuses on applying Group-DRO during the finetuning phase in order to improve WER for poor performing groups while also lowering the overall loss seen from our baseline experiments. We update the weights of the model $\boldsymbol{\theta}$ using stochastic gradient descent and the probability distribution for each group $\boldsymbol{q'}$ in following manner:

Input: Step sizes η_q, η_θ; P_g for each $g \in \mathcal{G}$
Initialize $\theta^{(0)}$ and $q^{(0)}$
for $t = 1, \ldots, T$ **do**
  $g \sim \text{Uniform}(1, \ldots,)$
  $x, y \sim P_g$
  $q' \leftarrow q^{(t-1)}$; $q'_g \leftarrow q'_g \exp(\eta_q \ell(\theta^{(t-1)}; (x, y)))$
  $q^{(t)} \leftarrow q' / \sum_{g'} q'_{g'}$
  $\theta^{(t)} \leftarrow \theta^{(t-1)} - \eta_\theta q_g^{(t)} \nabla \ell(\theta^{(t-1)}; (x, y))$
end

Algorithm 1: Group-DRO optimization algorithm.

The standard ML-SUPERB dataset contains only audio as well as the reference text transcription, thus during data preprocessing for our experiments, we append a **group index** field identifier which is used to update the correct group probabilities during Group-DRO finetuning.

5 Experiments

5.1 Baseline Experiments

We first evaluated Whisper-Small's speech recognition capabilities across low-resource languages from the Bantu family of languages including: Swahili, Amharic, Swati, Xhosa, Tswana, and others using Word Error Rate (WER) as our primary evaluation metric. We also include WER results on Dutch and English (the two languages with the most training examples in ML-SUPERB) for reference. Our baseline results are shown in Table 1.

Whisper-Small		
Language	Whisper Small	Low-Resource
Basaa	1.54	✓
Kinyarwanda	4.70	✓
Ganda	1.35	✓
Northern Sotho	1.73	✓
Chichewa	1.58	✓
Swati	1.89	✓
Swahili	1.36	✓
Tswana	3.26	✓
Venda	0.93	✓
Xhosa	3.54	✓
Zulu	2.31	✓
Dutch	0.21	
English	0.13	

Table 1: Baseline WER Across Bantu Family Languages and more.

We use language specific processors for inference on Whisper where possible, and use the Swahili tokenizer provided by Whisper for Bantu languages that do not have their own processor. We can see that the pretrained Whisper baselines have fairly high average WERs, especially in comparison to high resource languages like English and Dutch.

As an additional baseline, we finetune Whisper-Tiny on the 1 hour training set of 3 selected Bantu

Model	Language WER												
	bas	kin	lug	nso	nya	ssw	swa	tsn	ven	xho	zul	**avg**	**max**
whisper-tiny	1.23	2.66	1.22	1.24	2.18	1.43	2.22	2.39	1.15	3.13	2.78	2.08	3.13
+ ERM	0.70	1.47	0.76	0.48	0.81	0.62	0.63	0.51	0.61	0.62	0.65	0.66	1.47
+ GroupDRO	0.72	**1.12**	0.75	0.58	0.75	0.72	0.66	0.71	0.65	0.85	0.76	0.73	1.12
whisper-base	1.46	2.29	1.15	1.18	1.40	2.62	2.31	2.23	1.20	1.98	1.80	1.86	2.62
+ ERM	0.58	0.93	0.74	0.38	0.73	0.51	0.53	0.39	0.49	0.52	0.53	0.52	0.93
+ GroupDRO	0.62	1.13	0.71	0.44	0.68	0.57	0.58	0.51	0.59	0.58	0.57	0.59	1.13
whisper-small	1.54	4.70	1.36	1.73	1.59	1.90	1.37	3.26	0.93	3.55	2.31	2.32	4.70
+ ERM	0.56	0.89	0.68	0.33	0.61	0.43	0.42	0.26	0.43	0.41	0.41	0.44	0.89
+ GroupDRO	0.52	0.85	0.68	0.33	0.49	0.45	0.45	0.35	0.45	0.47	0.47	0.47	0.85

Table 2: Group-DRO improves worst case group performance compared to ERM at the cost of Average WER

Model	Source WER										
	cv	fl	nchlt	vf	mls	vp	swc	mai	lad	**avg**	**max**
whisper-tiny + ERM	0.42	0.28	0.24	0.15	0.30	0.20	0.33	0.16	0.08	0.24	0.42
whisper-tiny + Group-DRO	0.39	0.27	0.26	0.16	0.29	0.20	0.34	0.14	0.11	0.24	0.39
whisper-base + ERM	0.34	0.23	0.22	0.11	0.25	0.16	0.30	0.13	0.15	0.21	0.34
whisper-base + Group-DRO	0.33	0.20	0.22	0.11	0.24	0.17	0.29	0.13	0.20	0.21	0.33
whisper-small + ERM	0.31	0.19	0.20	0.11	0.20	0.15	0.26	0.17	0.11	0.19	0.31
whisper-small + Group-DRO	0.28	0.19	0.19	0.09	0.20	0.14	0.24	0.11	0.09	0.17	0.28

Table 3: Word Error Rates (WER) across different sources: Group-DRO obtains lower average and max WER compared to ERM across all models

languages - Swahili, Swati, and Xhosa. We show our results in Table 4. We train for 300 steps, with batch size 64 and learning rate 10^{-4}, and find that we are able to significantly improve performance simultaneously across all three languages, despite mixing training examples, and using the Swahili tokenizer for all three.

Whisper-Tiny Fine Tuned	
Language	WER
Swahili	0.58
Swati	0.57
Xhosa	0.66

Table 4: Whisper-tiny Baseline trained on **swa**,**ssw**, and **xho**

5.2 Group-DRO/ERM Finetuning

For our core experiments we finetune the three Whisper models using two optimization strategies: ERM and Group-DRO in order to compare how they affect WER. Our groups that we choose are both language dependent (Bantu family) as well as source dependent (English data from various sources of ML-SUPERB benchmark).

Our training setup that we use for our experiment runs are as follows: we perform 1000 training steps, use stochastic gradient descent with cross entropy loss as our learning objective, use a linear learning rate scheduler for stabilized convergence, and a batch size of 64.

In our Bantu language family experiments, we obtain results for 3 Whisper models using 3 different training strategies: zero-shot learning, ERM optimization, and Group-DRO optimization. Table 2 showcases the results we achieve. We see that models evaluated in the zero-shot setting without any finetuning lead to a high average WERs, demonstrating that the models struggle significantly without further finetuning on lower-resource languages, even for languages that were included in the pretraining set like Swahili (swa). Finetuning with ERM and Group-DRO optimization strategies lead to a $60\%+$ decrease in WER across all models and languages, showing that the models perform well at transcribing low-resource languages after applying both strategies. We find that for the Whisper-Tiny and Small models, the maximum group WER across Bantu languages is lower for Group-DRO compared to ERM, suggesting that Group-DRO's optimization approach has succeeded in reducing errors across worst-case language distributions compared to a standard ERM approach. More specifically, languages such as Kinyarwanda (kin), which has a top three WER prior to finetuning,

experiences a lower WER with Group-DRO compared to ERM. Thus, from our results, we observe that Group-DRO improves worst case group performance for some groups compared to ERM, with the tradeoff of a higher overall WER across languages which we discuss further in **Conclusion**.

For our experiments in which we organize groups by data source within ML-SUPERB (where all training data is English examples), we follow a similar process of obtaining results for 3 Whisper models using ERM and Group-DRO objectives. Table 3 summarizes the WERs we get across sources.

We find that Group-DRO consistently achieved the same or better WERs compared to ERM across all three models. Specifically, for the Whisper-Tiny model, the average WER remained the same (0.24) under both ERM and Group-DRO, but the maximum Group WER decreased from 0.42 to 0.39 with Group-DRO. This indicates that Group-DRO was more effective at handling the worst-performing data sources, thus improving overall robustness.

For the Whisper-Base model, the average WER was identical (0.21) for both objectives. However, the maximum group WER under Group-DRO (0.33) was slightly lower than under ERM (0.34). Likewise, the Whisper-Small model showed a similar trend. While the average WER improved slightly from 0.19 under ERM to 0.17 under Group-DRO, the maximum WER saw a more noticeable reduction, from 0.31 to 0.28. These slight improvements suggests that Group-DRO managed to reduce errors for the worst-performing groups while maintaining good overall performance across all groups.

6 Conclusion

We found through our methodology and experiments that finetuing with Group-DRO leads to significant improvements for worst-case performing groups and also lowers the overall average and maximum WER across groups compared to zero-shot learning. We compared the results of Group-DRO to the standard ERM optimization strategy and found tradeoffs between the average WER and maxiumum WER, with ERM in most cases leading to a lower average WER whereas Group-DRO led to a lower maximum WER in most cases.

Although a model with a lower average WER might seem more optimal, it's important to consider the use cases of an ASR system which need to be robust to many group distributions. For example, a finetuned ASR system optimized with an ERM objective might perform generally well across **most** of the low-resource groups it was finetuned on, but might perform noticeably worse for one group compared to the others, thus leading to the inequities discussed in **Introduction**. We argue that despite potentially yielding a higher average WER under certain circumstances (contingent upon group division), Group-DRO offers a more favorable outcome by mitigating the maximum group WER. This outcome is preferable as it ensures a more equitable distribution of performance across various groups which is especially important in real-world applications.

Our results show that Group-DRO can have impacts beyond the applications explored in the original paper (object recognition and natural language inference) and can be applied to ASR systems for the purpose of improving their robustness across low and high-resource groups. In our paper, we categorized these groups by language and data source, however, this is an avenue of future work that ought to be further explored. Different groupings might include categorizing by accents from a particular language, age, etc in order to further improve the robustness of ASR systems. This of course is dependent on the desired application for the system.

Lastly, we argue that distribution shift in ASR is a research field that ought to be further explored in order to understand how languages, dialects, groups, and other speech characteristics change over time, and how to develop robust systems that are capable of keeping up with these shifts. As we have noted in our paper, understanding distribution shift can help in designing ASR systems that are more inclusive and fair, reducing biases that can arise from underrepresented groups in the training data. It can also facilitate the creation of adaptive learning algorithms that continuously improve as they are exposed to new data, thus enhancing the long-term effectiveness and reliability of ASR technologies. Furthermore, addressing distribution shift is essential for the deployment of ASR models in critical applications where accuracy and reliability are of upmost importance, such as: healthcare, legal transcription, and emergency response systems. Thus, this area of research holds significant potential for advancing the capabilities and applications of ASR systems in the real world.

7 Contributions

JM and ET were both involved in model selection, gathering baseline results, experiment planning, integrating Group-DRO optimization into the codebase, and writing the final paper/poster. ET was responsible for data exploration and running experiments. JM was responsible for experiment analysis.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

John C Duchi, Peter W Glynn, and Hongseok Namkoong. 2021. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2018. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Jiatong Shi, Dan Berrebbi, William Chen, Ho-Lam Chung, En-Pei Hu, Wei Ping Huang, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung-yi Lee, et al. 2023. Ml-superb: Multilingual speech universal performance benchmark. *arXiv preprint arXiv:2305.10615*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.