Assessing and Mitigating Speech Model Biases via Pattern Mining

Alkis Koudounas
alkis.koudounas@polito.it
Politecnico di Torino
Turin, Italy

Eliana Pastor
eliana.pastor@polito.it
Politecnico di Torino
Turin, Italy

Elena Baralis
elena.baralis@polito.it
Politecnico di Torino
Turin, Italy

ABSTRACT

Speech models are often evaluated based on overall performance or predefined subgroups, which can miss valuable insights from more detailed subgroup analyses. Yet, identifying interpretable subgroups in raw speech data poses inherent challenges. To address this, we enhance speech data with metadata such as speaker demographics (gender, age), task-related features (intent, emotion), and signal-related metadata (speaking rate, number of pauses). This enriched data allows us to identify human-understandable subgroups where speech model performance deviates significantly from the average. Our approach is task-, model-, and dataset-agnostic, enabling us to identify performance disparities both within and across different models. We validate our method on three tasks (intent classification, automatic speech recognition, and emotion recognition), four datasets, and three models of varying sizes, providing nuanced insights into model assessments. We further leverage this approach to guide a mitigation strategy for improved and fairer models. The results demonstrate that our approach leads to performance improvements and significant reductions in performance disparities compared to random-, and clustering-based techniques.

KEYWORDS

Model bias analysis, Bias mitigation, Speech models, Subgroups

1 INTRODUCTION

Intelligent systems with speech recognition and comprehension capabilities are increasingly common. However, evaluations often focus on overall performance, neglecting disparities across different groups. Recent studies have highlighted issues of model bias and unequal treatment across data subgroups [2–4, 6, 8, 12]. A data subgroup is a subset of instances sharing similar characteristics or attribute values (e.g., utterances by female speakers). Previous approaches typically focused on predefined subgroups based on known attributes of interest, targeting biases in specific demographic traits such as skin tone [4], ethnicity [8], or combinations of metadata like gender, age, and accents [3]. However, these categorizations require human expertise and limit the exploration of unanticipated yet significant subgroups. We propose an automated method to identify critical subgroups to address these limitations. Unlike existing clustering-based techniques [2, 12], our approach allows intersectional analysis, exploring the combined impacts of multiple attributes. By combining metadata like speaker gender and features like speaking rate, we can identify interpretable subgroups.

Research questions. This study investigates bias in speech model performance across data subgroups via pattern mining. We automatically identify combinations of metadata values that exhibit the highest: (i) *intra-model performance gaps*, indicating significant performance differences between the overall dataset and specific data subgroups, and (ii) *cross-model performance gaps*, signifying notable differences in subgroup performance among different models. We leverage this interpretable identification of critical subgroups to design a data acquisition strategy to enhance performance and mitigate model biases. Therefore, this work addresses the following research questions (RQs): **(RQ1)** "How can we automatically identify and characterize the most critical subgroups for a speech model?", **(RQ2)** "How do model size and architecture impact subgroup performance?", and **(RQ3)** " How effectively can we mitigate these model biases through a subgroup-guided data acquisition?".

2 METHODOLOGY

We analyze speech model behavior by slicing data into interpretable subgroups. We define *interpretable metadata* as attributes understandable by humans, e.g., speaker age or gender or utterance noise level. For instance, *"old men in noisy scenarios"* is an interpretable subgroup. Identifying interpretable subgroups in raw speech data poses intrinsic challenges. To overcome this issue, we enrich speech data with interpretable metadata, including: (i) *speaker demographics* like gender or age, (ii) *task-specific features*, like intent or emotion associated with an utterance, (iii) *recording and speaking conditions*, such as noise level, speaking rate, and duration of pauses.

Items and Itemsets. Let D represent our dataset and A denote its metadata attribute set. An *item* represents an attribute equality $a = v$, where a is an attribute in A, and v is its value. Examples of items include $gender = male$ and $age \in [41 - 65]$. Note that continuous-valued attributes are discretized. A *subgroup* corresponding to an item denotes the dataset portion satisfying it. Items facilitate the selection of data subsets based on single attributes, while *itemsets* allow slicing across multiple attributes. An itemset I comprises zero or more items, each including a different attribute. For instance, an itemset like $\{gender = female, age \in [22, 40]\}$ defines a subgroup. The itemset *support* denotes the fraction of the dataset it covers. For instance, an itemset with support 0.02 represents 2% of the dataset. An itemset is *frequent* if its support exceeds a minimum threshold.

Intra and cross-model performance gaps We aim to identify subgroups with performance disparities relative to the overall dataset. We use DivExplorer [9] to extract all *frequent* itemsets above a certain support threshold via frequent pattern mining techniques. While the number of subgroups can increase exponentially with more attributes, many itemsets may have minimal support and thus be less relevant. Additionally, low-support subgroups can suffer from statistical fluctuations. Therefore, we focus only on frequent subgroups that exceed the support threshold to ensure meaningful analysis. We use the concept of subgroup divergence (i.e., intra-model performance gap) from [9]. Let f be a generic statistic for a speech task. For a model M and subgroup I, $f(I, M)$ represents the average statistic (e.g., accuracy, error rate) for the model on the subgroup. The divergence of itemset I for model M is defined

Table 1: *Top, RQ1.* Intra-model f gap for the most negatively ($\mathtt{I}^-$) and positively ($\mathtt{I}^+$) divergent subgroups. *Bottom, RQ2.* Cross-model f gap when scaling up wav2vec 2.0. (↑) and (↓) denote the highest performance improvement and decrease, respectively.

Subgroups	*Sup_train*	*Sup_test*	f	Δ_f	t
$\mathtt{I}^-$: {*age=22-40, gender=male, location=none, speaking rate=high, tot silence=high*}	0.03	0.04	60.50	-31.22	7.05
$\mathtt{I}^+$: {*age=22-40, location=washroom, speaking rate=low, trimmed duration=high*}	0.03	0.03	100.0	8.28	9.74
Subgroups	*Sup*	gap_f	$f_{\mathtt{w2v2-b}}$	$f_{\mathtt{w2v2-l}}$	t
↑ : {*action=increase, location=none, tot duration=low, trimmed speaking rate=low, trimmed duration=low*}	0.03	22.69	75.63	98.32	5.37
↓ : {*action=activate, gender=male, speaking rate=low*}	0.03	-20.97	96.77	75.81	4.92

Table 2: *RQ3. Original* wav2vec 2.0 training, *random* and *clustering*-based approaches, and *ours* strategy. Best results in bold.

K	*Approach*	*#samples*	*Accuracy*	*F1 Macro*	Δ^-_{max}	Δ^-_{avg-10}	Δ^-_{avg-20}	Δ^-_{avg-50}	$\lvert\Delta_{avg-all}\rvert$
-	original	18506	91.58 ± 0.08	86.34 ± 0.13	-70.09 ± 0.26	-70.09 ± 0.26	-65.73 ± 0.49	-53.31 ± 0.19	1.06 ± 0.07
	random	+226	92.56 ± 0.44	90.25 ± 0.60	-52.20 ± 2.57	-51.11 ± 2.19	-46.61 ± 1.34	-43.98 ± 0.68	0.97 ± 0.02
2	clustering	+226	89.77 ± 0.88	87.02 ± 0.15	-47.37 ± 0.42	-47.34 ± 0.42	-47.23 ± 0.43	-46.75 ± 0.91	0.94 ± 0.04
	ours	+226	**96.55 ± 0.08**	**94.71 ± 0.12**	**-40.60 ± 0.35**	**-40.28 ± 0.36**	**-38.08 ± 0.36**	**-32.72 ± 0.28**	0.81 ± 0.03
-	all data	+4606	93.42 ± 0.17	93.11 ± 0.17	-53.18 ± 0.15	-50.89 ± 0.09	-45.61 ± 0.14	-40.37 ± 0.16	**0.37 ± 0.01**

as the difference between the model's performance on I and its performance on the entire dataset: $\Delta_f(I, M) = f(I, M) - f(\emptyset, M)$. A higher absolute divergence indicates a more significant variation in subgroup performance compared to the overall dataset. To compare models, we introduce the cross-model performance gap, which measures the performance difference between two models on a specific subgroup. This gap helps compare models with different sizes, architectures, or pre-training objectives. For models M_1 and M_2, the performance gap for itemset I is the change in performance on I when replacing M_1 with M_2: $gap_f(I, M_1, M_2) = f(I, M_2) - f(I, M_1)$. The definitions of intra- and cross-model gaps apply to any speech model and task, allowing for subgroup performance assessment on any dataset with metadata. This methodology is task-, model-, and dataset-agnostic. We use Welch's t-test to evaluate the statistical significance of the differences in the statistic f between (i) the subgroup I and the dataset D, and (ii) the two models M_1 and M_2.

Subgroup-guided Data Acquisition After evaluating a speech model's performance, we aim to improve it overall and across different subpopulations. We identify critical subgroups with negative divergence, indicating challenging scenarios for the model. We use a pruning procedure from [9] to reduce redundancy. When two subgroups, I_a and I_b, have similar divergences but I_b includes an additional metadata condition, we retain only the more general subgroup, I_a. We then prioritize data acquisition for the top-K critical subgroups with the highest negative divergence and retrain the model with additional data from these subgroups. The parameter K controls the extent of data acquisition. More details in [5].

3 RESULTS AND DISCUSSION

We evaluate our method by (i) analyzing its ability to identify sources of errors, (ii) examining the influence of factors like model size and architecture on subgroup-level performance, and (iii) assessing the effectiveness of subgroup-guided data acquisition on bias mitigation. Further details can be found in [5–7].

RQ1: Model understanding. We focus on wav2vec 2.0 base model [1] performance, i.e., accuracy. Table 1(top) highlights the subgroups with the largest divergence for the FSC dataset in the intent classification task. Negative divergence signifies below-average accuracy, while positive divergence indicates above-average accuracy. These divergence values are statistically significant (with $t > 2$, per Siegel's rule of thumb [11]). The model performs worst for the subgroup of male speakers aged 22-40, unspecified location, high speaking rate, and high total silence (divergence $\Delta_f = -31.2\%$). Conversely, it performs best for the subgroup of speakers aged 22-40, with low speaking rate, long duration, and "washroom" as the target location, correctly predicting all utterances. For further analyses on other datasets and models, please see [7].

RQ2: Model comparison. We compare model performances overall and across subgroups to identify which populations benefit most from model changes. Specifically, we analyze the impact of model size. Larger models generally show higher overall accuracy, and [10] claims that they are also fairer. However, subgroup performance depends on the dataset/task. Table 1(bottom) summarizes performance changes for FSC when scaling up wav2vec 2.0 size. We find varying subgroup impacts, with some groups benefiting more than others. However, over 30% of explored subgroups experience decreased performance with larger model sizes. For other datasets and factors (e.g., architecture), please refer to [7].

RQ3: Subgroup-guided data acquisition. We utilize critical subgroups to guide targeted data acquisition, focusing on $k = 2$, that is, the two highest-divergent subgroups, for FSC. Further details on different k values can be found in [5]. We partition the dataset into training, held-out, validation, and test sets. After identifying critical subgroups using the validation set, we acquire data samples from the held-out set and retrain the model. Evaluation on the test set consistently shows superior performance compared to baseline methods like indiscriminate random and clustering-guided acquisition (Table 2). Selecting only the top 2 critical subgroups leads to significant performance improvements, achieving the best F1 score and accuracy performance, as well as the lowest maximum divergence (Δ^-_{max}) and the lowest average divergence for the top-10 (Δ^-_{avg-10}), 20 (Δ^-_{avg-20}), and 50 (Δ^-_{avg-50}) subgroups with the highest negative divergence. Targeted data acquisition effectively mitigates performance disparities and improves model robustness.

REFERENCES

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, Vol. 33. 12449–12460.

[2] 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 *Proc. Interspeech 2022*. 1268–1272. https://doi.org/10.21437/Interspeech.2022-10816

[3] Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122* (2021).

[4] 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. *Proc. of the National Academy of Sciences* 117, 14 (2020), 7684–7689.

[5] Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Luca de Alfaro, and Elena Baralis. 2024. Prioritizing Data Acquisition for end-to-end Speech Model Improvement. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7000–7004. https://doi.org/10.1109/ICASSP48485.2024.10446326

[6] Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Luca Cagliero, Luca de Alfaro, Elena Baralis, and Daniele Amberti. 2023. Exploring Subgroup Performance in End-to-End Speech Models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095284

[7] Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Elisa Reale, Luca Cagliero, Sandro Cumani, Luca de Alfaro, Elena Baralis, and Daniele Amberti. 2024. Towards Comprehensive Subgroup Performance Analysis in Speech Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), 1468–1480. https://doi.org/10.1109/TASLP.2024.3363447

[8] Li-Fang Lai and Nicole Holliday. 2023. Exploring Sources of Racial Bias in Automatic Speech Recognition through the Lens of Rhythmic Variation. In *Proc. INTERSPEECH 2023*. 1284–1288. https://doi.org/10.21437/Interspeech.2023-159

[9] Eliana Pastor, Luca de Alfaro, and Elena Baralis. 2021. Looking for Trouble: Analyzing Classifier Behavior via Pattern Divergence. In *Proceedings of the 2021 International Conference on Management of Data* (Virtual Event, China) *(SIGMOD '21)*. ACM, 1400–1412. https://doi.org/10.1145/3448016.3457284

[10] Yi Sheng, Junhuan Yang, Yawen Wu, Kevin Mao, Yiyu Shi, Jingtong Hu, Weiwen Jiang, and Lei Yang. 2022. The Larger The Fairer? Small Neural Networks Can Achieve Fairness for Edge Devices. *arXiv preprint arXiv:2202.11317* (2022).

[11] Andrew F. Siegel. 2012. Chapter 10 - Hypothesis Testing: Deciding between Reality and Coincidence. In *Practical Business Statistics (Sixth Edition)* (sixth edition ed.), Andrew F. Siegel (Ed.). Springer Science & Business Media, 249–287. https://doi.org/10.1016/B978-0-12-385208-3.00010-9

[12] 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)*. IEEE, 1–5.