Assessing Speech Model Performance: A Subgroup Perspective

Assessing Speech Model Performance: A Subgroup Perspective

Alkis Koudounas[1,*], Eliana Pastor[1] and Elena Baralis[1]

[1]Politecnico di Torino, Turin, Italy

Abstract

Spoken language understanding (SLU) models are commonly evaluated based on overall performance or predefined subgroups, often overlooking the potential insights gained from more comprehensive subgroup analyses. Conducting a more thorough analysis at the subgroup level can reveal valuable insights into the variations in speech system performance across different subgroups. Yet, identifying interpretable subgroups in raw speech data poses inherent challenges.

To overcome these issues, we enrich speech data with metadata from various domains. We consider, when available, speaker demographics like gender, age, and origin country. We also incorporate task-related features, such as a specific intent or emotion associated with an utterance. Finally, we extract signal-related metadata, including speaking rate, signal-to-noise ratio, number of words, and number of pauses. Including these features, extracted directly from the raw signal, is crucial in capturing fine-grained nuances that may impact model performance. By combining these metadata, we identify human-understandable subgroups in which speech models exhibit performance significantly better or worse than the average.

Our approach is task-, model-, and dataset-agnostic. It enables the identification of intra- and cross-model performance gaps, highlighting disparities among different models. We validate our methodology across three tasks (intent classification, automatic speech recognition, and emotion recognition), three datasets, and one speech model with different sizes, providing nuanced insights into model assessments. We further propose leveraging this approach to guide a data acquisition strategy for improved and fairer models. The experimental results demonstrate that our approach leads to substantial performance improvements and significant reductions in performance disparities, all achieved with reduced data and costs compared to random and clustering-based acquisition techniques.

Keywords

Subgroup identification, Model bias analysis, Bias mitigation, Speech representation, E2E-SLU models

1. Introduction

Intelligent systems with speech recognition, transcription, and comprehension capabilities are increasingly common across various domains, including virtual assistants [1, 2], customer service [3, 4], and healthcare [5, 6]. However, current evaluation paradigms for these systems predominantly focus on aggregate performance metrics, overlooking potential disparities across different groups [7, 8, 9]. Furthermore, the proliferation of large pre-trained neural models using self-supervised learning poses challenges for interpretability and identification of performance disparities through conventional methodologies [10, 11]. These issues highlight the need for a

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23–26, 2024, Villasimius, Sardinia, Italy

*Corresponding author.

alkis.koudounas@polito.it (A. Koudounas); eliana.pastor@polito.it (E. Pastor); elena.baralis@polito.it (E. Baralis)

0000-0003-4386-0409 (A. Koudounas); 0000-0002-3664-4137 (E. Pastor); 0000-0001-9231-467X (E. Baralis)

comprehensive evaluation framework that captures subgroup-level effects to enable responsible assessment of speech technologies, identifying and mitigating unintended harms.

Recent literature has highlighted issues of model bias and unequal treatment across data subgroups [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. A data subgroup refers to a subset of instances demonstrating similar characteristics within the latent space or common attribute values (e.g., utterances spoken by female speakers). Previous approaches have typically focused on predefined subgroups based on protected attributes or features of interest known *a priori*. Specifically, these works targeted identifying bias within specific demographic traits, such as skin tone [12], ethnicity [16], or combinations of metadata, such as demographics and geolocation [15], as well as gender, age, and accents [13] or gender, age, skin tones [14]. However, such categorizations often necessitate human expertise and preclude the exploration of unanticipated yet significant subgroups.

In this work, we propose an automated method for identifying critical subgroups to address these limitations. Unlike existing clustering-based speaker embedding techniques [15, 18], our approach facilitates intersectional analysis, enabling us to explore the combined impacts of multiple attributes. Speech data frequently includes additional metadata about the speaker (e.g., the gender) or task (e.g., the emotion associated with a sentence). Other features as speaking rate, signal-to-noise ratio, and number of words, can be extracted from the audio or transcripts. The latter are essential for capturing narrow nuances that could significantly affect model performance. By combining such metadata values, we can identify interpretable data subgroups.

Research questions. This study investigates bias in speech model performance across data subgroups, mainly focusing on spoken language understanding (SLU). 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. Our approach enables the identification of data subgroups where a model exhibits lower performance compared to the overall behavior. We leverage this interpretable identification of critical subgroups for a targeted 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 an SLU model?", **(RQ2)** "How does model size or architecture impact subgroup performance?", and **(RQ3)** " How does adopting a subgroup-guided data acquisition strategy influence the overall model and subgroup performance compared to an indiscriminate approach?".

Our approach. We introduce a novel task-, model-, and dataset-agnostic methodology for automating the characterization and comparison of data subgroups induced by metadata attributes. We identify all "frequent subgroups," i.e., those exceeding a certain support threshold (e.g., at least 0.1% of the dataset), that exhibit maximal disparities in intra- and cross-model performance. We provide end-users with interpretable representations of such critical subgroups within a given speech task and model and further use this information to mitigate model inner biases.

The primary contributions of this work are: (i) a novel framework for analyzing SLU models by identifying subgroups exhibiting large performance gaps; (ii) insights into the effects of model size at the subgroup level; and (iii) a subgroup-guided targeted data acquisition approach

to enhance overall and across subgroups model performance.

We conduct comprehensive experiments across three speech tasks (Automatic Speech Recognition (ASR), Intent Classification (IC), Emotion Recognition (ER)), three datasets (LibriSpeech [24], FSC [25], and IEMOCAP [26]), and for the transformer-based speech model wav2vec 2.0 [27]. Our experimental results demonstrate that our subgroup-level analysis reveals distinctive performance patterns in data subpopulations. We further show that our subgroup-guided acquisition approach consistently improves performance both overall and on subgroups compared to an indiscriminate strategy, even when acquiring a subset of the data.

2. Methodology

Our approach examines model performance at the subgroup level, where a *subgroup* is defined as a subset of the data characterized by specific metadata values, and denoted as itemset. This metadata covers mixed factors, including speaker traits (e.g., gender, age), speech features (e.g., speaking rate, number of pauses), and task-specific attributes (e.g., intents, labels). For instance, the subgroup *{gender=male, age ∈ [41-65]}* signifies utterances from male speakers aged 41 to 65.

Our analysis of subgroup behavior leverages two key concepts: intra-model divergence and cross-model performance gap. The former indicates the disparity in model performance between a subgroup and the entire dataset, revealing subgroups associated with performance variations, be it below-average, above-average, or equivalent. We will also leverage this aspect to guide the data acquisition strategy. Conversely, the latter quantifies the performance differences between two models on the same subgroup, facilitating comparative assessments at the subgroup level.

2.1. Itemsets through interpretable metadata

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.

Metadata Description. Identifying interpretable subgroups in raw speech data poses intrinsic challenges. To overcome this issue, we enrich speech data with interpretable metadata from various domains, providing a human-understandable description of utterances. They can be inherent to the dataset or derived from utterances/transcriptions. Examples of such metadata attributes include: (i) *speaker demographics* like gender or age, (ii) *task-specific features*, like intent or emotion associated with an utterance, (iii) *recording conditions*, such as environment type and noise level, and (iv) *speech features*, such as speaking rate and duration of silences.

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. We only focus on discretized attributes, thus continuous-valued attributes are discretized before applying our techniques. Examples of items include $gender = male$ and $age \in [41 - 65]$, if *gender* and *age* are attributes. A *subgroup* corresponding to an item denotes the dataset portion satisfying it. We ensure that subgroups form a dataset partition for each attribute. For example, the age ranges must not overlap within the *age* attribute, and collectively, they must cover all potential age ranges.

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 based on the gender and age attributes. We define data subgroups via itemsets, enabling an interpretable subgroup definition. The *support* of an itemset denotes the fraction of the dataset it covers. For instance, an itemset with support 0.02 represents 2% of the dataset. The empty itemset ($\emptyset$) corresponds to the entire dataset and has a support of 1. An itemset is *frequent* if its support exceeds a minimum threshold (u).

2.2. Intra and cross-model performance gaps

We aim to identify subgroups exhibiting performance disparities compared to the overall dataset. We rely on DivExplorer [22, 28] to extract all frequent itemsets above a specified support threshold. While subgroups grow exponentially with the number of attributes, many extracted itemsets may have minimal or zero support, making them less relevant for subgroup performance analysis. Performance statistics for subgroups with low support may also suffer from statistical fluctuations. Therefore, to ensure operational significance, we only focus on the subgroups surpassing a given threshold (e.g., comprising at least 0.1% of the dataset), called frequent itemsets, which tend to be more limited.

We employ the concept of subgroup divergence (i.e., intra-model performance gap) as introduced in [22]. It quantifies the difference in performance between a subgroup and the entire dataset. Let f represent a generic statistic for a downstream SLU task. For a model M and a subgroup (i.e., itemset) I, $f(I, M)$ denotes the average statistic value (e.g., accuracy, error rate) of the model on the subgroup. We define the divergence of itemset I for model M as the difference between the model performance over I and the performance over the entire dataset:

$$\Delta_f(I, M) = f(I, M) - f(\emptyset, M) \tag{1}$$

A higher divergence (in absolute terms) indicates a more significant variation in subgroup performance compared to the overall dataset.

Assessing performance discrepancies at the subgroup level is also crucial for model comparison. We introduce the concept of cross-model performance gap, which measures the performance difference between two models on a specific subgroup. This gap could be used to compare different models, characterized by different size, architecture, or pre-training objective. Specifically, given two models M_1 and M_2, the performance gap from model M_1 to model M_2 for the itemset I is defined as the change in performance on I obtained by replacing M_1 with M_2:

$$gap_f(I, M_1, M_2) = f(I, M_2) - f(I, M_1) \tag{2}$$

The definitions of intra- and cross-model gaps apply to generic SLU models for any task, enabling assessment of subgroup performance for a given dataset annotated via metadata. This methodology thus remains task-, model-, and dataset-agnostic. To evaluate the statistical significance, we employ Welch's t-test to test the hypothesis that the means of the statistic f are equal for (i) the subgroup I and the entire population D, and (ii) the two models M_1 and M_2.

Table 1
RQ1. Intra-model performance gap in the f measure (accuracy, or WER for LibriSpeech) for the most negatively (I^-) and positively (I^+) divergent subgroups compared to overall test performance.

Dataset	*Subgroups*	*Sup_train*	*Sup_test*	f	Δ_f	t
FSC	I^-: {*age=22-40, gender=male, location=none, speaking rate=high, tot silence=high*}	0.03	0.04	60.50	-31.22	7.05
	I^+: {*age=22-40, location=washroom, speaking rate=low, trimmed duration=high*}	0.03	0.03	100.0	8.28	9.74
IEMO	I^-: {*label=happy, activation=low*}	0.03	0.03	44.74	-29.92	7.37
	I^+: {*label=sad, valence=low, tot silence=low, trimmed duration=high*}	0.03	0.03	98.57	23.92	17.01
LS	I^-: {*gender=female, trimmed speaking rate=high, trimmed duration=low, num pauses=low*}	0.05	0.03	17.30	11.24	4.16
	I^+: {*gender=female, speaking rate=low, trimmed speaking rate=low, num pauses=low, tot duration=medium*}	0.03	0.03	3.27	-2.79	5.57

2.3. Local contribution through Shapley values

After identifying itemsets exhibiting significant divergence or gap, we seek to understand the contribution of each item to these metrics. We employ game theory concepts to provide local insights into subgroup behavior.

The *local* contribution quantifies the role of each item within an itemset in influencing its divergence or gap, using Shapley values. In this framework, items within an itemset are akin to team members, and the divergence or gap metric represents the team's total score. Specifically, for an item i within itemset I and a metric of interest $g(I)$, i.e., divergence or gap, the Shapley value $s_g(i, I)$ measures how much i contributes to $g(I)$, with $\sum_{i \in I} s_g(i, I) = g(I)$. More details on this local as well as the global contribution can be found in [17, 22, 29].

2.4. Subgroup-guided Data Acquisition

After evaluating the performance of a given speech model, our objective is to improve it both overall and across different subpopulations. We identify the critical subgroups (i.e., itemsets) characterized by negative divergence, representing challenging scenarios for the model. We implement a pruning procedure to eliminate redundancy among such subgroups, following [22]. Specifically, when encountering two subgroups, I_a and I_b, where I_b includes I_a along with an additional metadata condition, we retain only the more general subgroup, I_a, if the absolute difference in their divergences is below a predefined threshold. This approach is based on the rationale that I_a adequately captures the divergence exhibited by I_b, as the extra metadata in I_b only marginally affects the divergence. Pruning the critical subgroups yields a more concise representation, forcing the data acquisition process to focus on the most pertinent attributes.

We prioritize data acquisition efforts on the top-K critical subgroups with the highest negative divergence in accuracy and retrain the model with additional data belonging to these subgroups. The parameter K allows us to control the data acquisition process and observe its impact on model performance overall and within subgroups. Further details can be found in [30].

3. Results and Discussion

We assess the effectiveness of our methodology by (i) analyzing its ability to identify sources of errors, (ii) examining the influence of factors such as model size, architecture, and pre-training

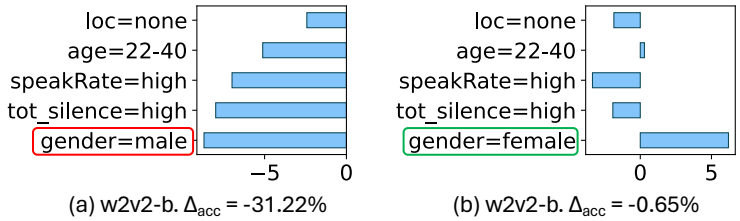


Figure 1: RQ1. Local contribution to accuracy divergence; FSC dataset, wav2vec 2.0 base.

Table 2
RQ2. Cross-model performance gap for f (WER for LIBRISPEECH, accuracy for the others) when scaling up wav2vec 2.0 size, from base (90 million parameters) to large (300 million parameters). (↑) denotes the highest performance improvement, (↓) indicates the largest decrease.

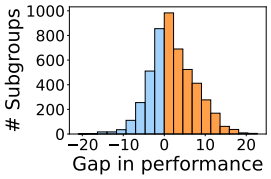
Dataset	*Subgroups*	*Sup*	gap_f	$f_{w2v2\text{-}b}$	$f_{w2v2\text{-}l}$	t
FSC	↑ {*action=increase, location=none, tot duration=low, trimmed speaking rate=low, trimmed duration=low*}	0.03	22.69	75.63	98.32	5.37
	↓ {*action=activate, gender=male, speaking rate=low*}	0.03	-20.97	96.77	75.81	4.92
IEMO	↑ {*label=happy, trimmed speaking rate=low*}	0.04	12.96	67.28	80.25	2.66
	↓ {*label=sad, trimmed speaking rate=low*}	0.03	-19.86	70.55	50.68	3.53
LS	↑ {*gender=female, num pauses=low, trimmed speaking rate=high, trimmed duration=low*}	0.03	-5.97	17.30	11.33	1.78
	↓ {*gender=male, num pauses=low, tot duration=low, trimmed speaking rate=high, trimmed duration=low*}	0.04	0.46	10.17	10.64	0.14

objective on subgroup-level performance, and (iii) evaluating the effect of using subgroup-level information to guide a data acquisition strategy in enhancing model performance and mitigating biases. Please refer to [17, 29, 30] for a complete set of the results.

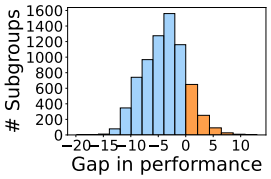
Metadata. We enrich the datasets with various metadata categories. We first incorporate demographic attributes of speakers where available, including gender, age, and country. We also consider unique metadata pertinent to each task if available, i.e., intent for FSC, and emotion and arousal labels for IEMOCAP. We finally extract from the raw signal utterance/transcription attributes such as silence duration (total and trimmed), word count, speaking rate (words per second), signal-to-noise ratio, and spectral flatness. The trimmed duration excludes initial and final pauses, while the total silence duration includes the entire utterance without any pauses. As the frequency and duration of intermediate pauses had little effect on model performance across all datasets, except for LIBRISPEECH, we chose to retain them for this dataset only.

Continuous attributes like speaking rate or utterance duration require discretization into fixed ranges. Using frequency-based discretization, we thus discretize this metadata into three ranges labeled as "low," "medium," and "high."

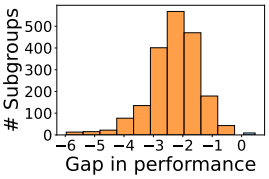
RQ1: Model understanding at the subgroup level. We focus on the performance of the wav2vec 2.0 base model [27] across all datasets. Table 1 shows the subgroups with the largest negative and positive divergence, indicating critical scenarios for each dataset. The divergence values associated with these subgroups are statistically significant (with $t > 2$, as per Siegel's rule of thumb [31]). For FSC and IEMOCAP, we evaluate model accuracy across various data subgroups, where higher accuracy indicates better performance. A negative divergence signifies accuracy below the average, while a positive divergence indicates above-average accuracy.



(a) **FSC.** Performance improvement for 63.75% of subgroups, decrease for 31.89% of them.



(c) **IEMOCAP.** Performance improvement for 11.21% of subgroups, decrease for 85.85% of them.



(d) **LibriSpeech.** Performance improvement for 99.25% of subgroups, decrease for 0.75% of them.

Figure 2: RQ2. Intra-model performance gap when scaling up wav2vec 2.0 model, from base to large.

For instance, for FSC, the wav2vec 2.0 base model exhibits its poorest performance for the subgroup characterized by speakers aged 22-40, male gender, no specified location, high speaking rate, and high total silence (Table 1, first block), with a divergence of $\Delta_{acc} = -31.2\%$. Analyzing sensitive attributes like gender is crucial, as evidenced by the significant impact observed. Specifically, female speakers achieve higher accuracy within the identified subgroup than males when all other metadata values remain constant. This trend is further confirmed by the Shapley values illustrated in Figure 1(a)-(b), where the male gender is associated with lower accuracy. In contrast, the female gender exhibits a positive impact.

Conversely, the analysis also reveals subgroups with above-average performance. For example, the model correctly predicts all utterances associated with the subgroup of speakers aged 22-40 with a low speaking rate, long duration, and "washroom" as the target location.

Similar assessments can be made for other datasets. For LibriSpeech, we study the Word Error Rate (WER); a positive WER divergence (i.e., higher than overall) signifies lower performance.

RQ2: Model comparison at the subgroup level. We compare different model performances at the overall and subgroup levels, detecting which subpopulations benefit the most from model changes. We analyze here how increasing the size of such models affects their performance at both levels. For changes in architecture and pre-training objective, please refer to [29].

Larger models tend to be more accurate overall, and [32] claims that larger models are also fairer. However, performance for specific subgroups is complex and depends on the dataset/task. We specifically examine how scaling up the wav2vec 2.0 model influences performance across datasets, with Table 2 summarizing the performance gap in terms of the highest performance improvement and decrease, and Figure 2 illustrating the distribution of this gap across subgroups.

While a larger model size enhances both overall and subgroup WER in the LibriSpeech dataset, it diminishes performance at both levels for IEMOCAP. We further reveal varying subgroup impacts on FSC, indicating that certain groups benefit more from a larger model size than others. Nonetheless, more than 30% of the explored subgroups decrease performance when scaling up the size. These findings emphasize the importance of analyzing subgroup-specific outcomes when evaluating the effectiveness of larger models.

RQ3: Subgroup-guided data acquisition. We use the identified critical subgroups to guide a targeted data acquisition to improve model performance and mitigate its biases. We discuss the results for FSC. Further outcomes on ITALIC [33], an IC dataset in Italian, can be found in [30].

We partition our dataset into training, held-out, validation, and test sets, employing an 80-20

Table 3
RQ3. Results for the *original* fine-tuning of wav2vec 2.0, two baselines (*random* and *clustering*-based) and *our* subgroup-aware strategy. Best results for each number of considered subgroups K are highlighted in bold, while best results overall are in light-blue.

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
2	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
	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**
3	random	+382	94.13 ± 0.58	91.51 ± 0.82	-52.99 ± 3.40	-51.92 ± 3.02	-49.39 ± 2.21	-45.98 ± 1.78	0.33 ± 0.04
	clustering	+382	90.03 ± 0.97	85.30 ± 0.94	-46.40 ± 0.36	-45.02 ± 0.33	-41.59 ± 0.28	-37.79 ± 0.16	0.81 ± 0.02
	ours	+382	93.62 ± 0.29	**92.96 ± 0.46**	**-42.23 ± 0.12**	**-42.21 ± 0.11**	**-41.48 ± 0.11**	**-33.61 ± 0.07**	**0.22 ± 0.02**
4	random	+422	92.64 ± 0.27	91.29 ± 0.21	-55.83 ± 2.11	-55.71 ± 2.04	-51.41 ± 1.74	-45.41 ± 1.74	0.39 ± 0.02
	clustering	+422	87.72 ± 0.71	83.42 ± 0.48	-47.59 ± 0.25	-46.98 ± 0.21	-45.69 ± 0.12	-43.98 ± 0.09	0.72 ± 0.03
	ours	+422	**95.16 ± 0.11**	**92.47 ± 0.22**	**-45.68 ± 0.24**	**-44.56 ± 0.25**	**-41.53 ± 0.24**	**-37.02 ± 0.20**	**0.15 ± 0.01**
5	random	+509	91.48 ± 0.55	90.27 ± 0.49	-54.82 ± 3.41	-54.75 ± 3.29	-54.69 ± 3.11	-51.12 ± 2.25	0.96 ± 0.08
	clustering	+509	91.44 ± 1.41	87.92 ± 1.38	-51.92 ± 0.19	-51.90 ± 0.24	-49.79 ± 0.18	-43.39 ± 0.11	0.45 ± 0.03
	ours	+509	**94.12 ± 0.13**	**92.57 ± 0.16**	**-49.33 ± 0.15**	**-49.29 ± 0.12**	**-48.11 ± 0.21**	**-39.01 ± 0.11**	**0.11 ± 0.02**
-	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

split for training and held-out data, respectively, while retaining the original validation and test sets. We first identify critical subgroups using the validation set, then acquire data samples from the held-out set, and retrain the model with these samples. Evaluation on the test set (Table 3) reveals consistently superior performance across overall and subgroup-level metrics, compared to baseline methods such as indiscriminate random and clustering-guided acquisition [15], where samples are selected from the acoustic embedding clusters with subpar performance.

Selecting only the top 2 critical subgroups leads to significant performance improvements at both overall and subgroup levels. Specifically, it achieves 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. While performance slightly lowers when increasing the number K of critical subgroups, it remains significantly better than the original model performance and the one obtained when adding all available data. The lowest average absolute divergence is found with $K = 5$ critical subgroups, indicating reduced performance disparities across subgroups.

Overall, these results underscore the effectiveness of targeted data acquisition in mitigating performance disparities and improving model robustness across diverse subgroups.

4. Conclusion

This study presents a novel methodology for evaluating spoken language understanding (SLU) system performance by analyzing model bias at the subgroup level. We enrich raw speech data by extracting metadata that include speaker demographics, task- and signal-related features to allow the definition of human-interpretable subgroups. By automating the detection of performance disparities within subgroups, our approach enhances error analysis, facilitates model comparison, and mitigates biases, thus improving overall performance. This versatile methodology demonstrates effectiveness across various speech tasks, datasets, and model sizes, offering insights into which subgroups benefit most from system enhancements and contributing to the development of more inclusive and effective speech technologies.

References

[1] R. Sarikaya, P. A. Crook, A. Marin, M. Jeong, J.-P. Robichaud, A. Celikyilmaz, Y.-B. Kim, A. Rochette, O. Z. Khan, X. Liu, et al., An overview of end-to-end language understanding and dialog management for personal digital assistants, in: 2016 ieee spoken language technology workshop (slt), IEEE, 2016, pp. 391–397.

[2] G. Terzopoulos, M. Satratzemi, Voice assistants and smart speakers in everyday life and in education, Informatics in Education 19 (2020) 473–490.

[3] M. Nuruzzaman, O. K. Hussain, A survey on chatbot implementation in customer service industry through deep neural networks, in: 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE), IEEE, 2018, pp. 54–61.

[4] S. Scheidt, Q. Chung, Making a case for speech analytics to improve customer service quality: Vision, implementation, and evaluation, International Journal of Information Management 45 (2019) 223–232. URL: https://www.sciencedirect.com/science/article/pii/S0268401217309441. doi:https://doi.org/10.1016/j.ijinfomgt.2018.01.002.

[5] S. Latif, J. Qadir, A. Qayyum, M. Usama, S. Younis, Speech technology for healthcare: Opportunities, challenges, and state of the art, IEEE Reviews in Biomedical Engineering (2020).

[6] M. La Quatra, L. Vaiani, A. Koudounas, L. Cagliero, P. Garza, E. Baralis, How much attention should we pay to mosquitoes?, in: Proceedings of the 30th ACM International Conference on Multimedia, MM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 7135–7139. URL: https://doi.org/10.1145/3503161.3551594. doi:10.1145/3503161.3551594.

[7] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally, et al., Hear: Holistic evaluation of audio representations, in: NeurIPS 2021 Competitions and Demonstrations Track, PMLR, 2022, pp. 125–145.

[8] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, H. yi Lee, SUPERB: Speech Processing Universal PERformance Benchmark, in: Proc. Interspeech 2021, 2021, pp. 1194–1198. doi:10.21437/Interspeech.2021-1775.

[9] M. La Quatra, A. Koudounas, L. Vaiani, E. Baralis, P. Garza, L. Cagliero, S. M. Siniscalchi, Benchmarking representations for speech, music, and acoustic events, in: 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), 2024.

[10] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 (2023).

[11] C. Singh, J. P. Inala, M. Galley, R. Caruana, J. Gao, Rethinking interpretability in the era of large language models, arXiv preprint arXiv:2402.01761 (2024).

[12] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, S. Goel, Racial disparities in automated speech recognition, Proc. of the National Academy of Sciences 117 (2020) 7684–7689.

[13] S. Feng, O. Kudina, B. M. Halpern, O. Scharenborg, Quantifying bias in automatic speech recognition, arXiv preprint arXiv:2103.15122 (2021).

[14] C. Liu, M. Picheny, L. Sarı, P. Chitkara, A. Xiao, X. Zhang, M. Chou, A. Alvarado, C. Hazirbas, Y. Saraf, Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 6162–6166.

[15] P. Dheram, M. Ramakrishnan, A. Raju, I.-F. Chen, B. King, K. Powell, M. Saboowala, K. Shetty, A. Stolcke, Toward fairness in speech recognition: Discovery and mitigation of performance disparities, in: Proc. Interspeech 2022, 2022, pp. 1268–1272. doi:10.21437/Interspeech.2022-10816.

[16] L.-F. Lai, N. Holliday, Exploring Sources of Racial Bias in Automatic Speech Recognition through the Lens of Rhythmic Variation, in: Proc. INTERSPEECH 2023, 2023, pp. 1284–1288. doi:10.21437/Interspeech.2023-159.

[17] A. Koudounas, E. Pastor, G. Attanasio, V. Mazzia, M. Giollo, T. Gueudre, L. Cagliero, L. de Alfaro, E. Baralis, D. Amberti, Exploring subgroup performance in end-to-end speech models, in: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5. doi:10.1109/ICASSP49357.2023.10095284.

[18] I.-E. Veliche, P. Fung, 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, 2023, pp. 1–5.

[19] A. Koudounas, F. Giobergia, Houston we have a divergence: A subgroup performance analysis of asr models, arXiv preprint arXiv:2404.07226 (2024).

[20] A. Koudounas, F. Giobergia, E. Baralis, Bad exoplanet! explaining degraded performance when reconstructing exoplanets atmospheric parameters, in: NeurIPS 2023 AI for Science Workshop, 2023.

[21] N. Shahbazi, Y. Lin, A. Asudeh, H. Jagadish, Representation bias in data: a survey on identification and resolution techniques, ACM Computing Surveys 55 (2023) 1–39.

[22] E. Pastor, L. de Alfaro, E. Baralis, Looking for trouble: Analyzing classifier behavior via pattern divergence, in: Proceedings of the 2021 International Conference on Management of Data, SIGMOD '21, ACM, 2021, p. 1400–1412. doi:10.1145/3448016.3457284.

[23] E. Pastor, E. Baralis, L. de Alfaro, A hierarchical approach to anomalous subgroup discovery, in: 39th IEEE International Conference on Data Engineering, ICDE 2023, IEEE, 2023, pp. 2647–2659. doi:10.1109/ICDE55515.2023.00203.

[24] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: An asr corpus based on public domain audio books, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206–5210. doi:10.1109/ICASSP.2015.7178964.

[25] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, Y. Bengio, Speech model pre-training for end-to-end spoken language understanding, in: Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, 2019, pp. 814–818.

[26] C. Busso, M. Bulut, C.-C. Lee, E. A. Kazemzadeh, E. M. Provost, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: interactive emotional dyadic motion capture database, Language Resources and Evaluation 42 (2008) 335–359.

[27] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 12449–12460.

[28] E. Pastor, A. Gavgavian, E. Baralis, L. de Alfaro, How divergent is your data?, Proc. VLDB

Endow. 14 (2021) 2835–2838. doi:10.14778/3476311.3476357.

[29] A. Koudounas, E. Pastor, G. Attanasio, V. Mazzia, M. Giollo, T. Gueudre, E. Reale, L. Cagliero, S. Cumani, L. de Alfaro, E. Baralis, D. Amberti, Towards comprehensive subgroup performance analysis in speech models, IEEE/ACM Transactions on Audio, Speech, and Language Processing 32 (2024) 1468–1480. doi:10.1109/TASLP.2024.3363447.

[30] A. Koudounas, E. Pastor, G. Attanasio, L. de Alfaro, E. Baralis, Prioritizing data acquisition for end-to-end speech model improvement, in: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 7000–7004. doi:10.1109/ICASSP48485.2024.10446326.

[31] A. F. Siegel, Chapter 10 - hypothesis testing: Deciding between reality and coincidence, in: A. F. Siegel (Ed.), Practical Business Statistics (Sixth Edition), sixth edition ed., Springer Science & Business Media, 2012, pp. 249–287. doi:10.1016/B978-0-12-385208-3.00010-9.

[32] Y. Sheng, J. Yang, Y. Wu, K. Mao, Y. Shi, J. Hu, W. Jiang, L. Yang, The larger the fairer? small neural networks can achieve fairness for edge devices, arXiv preprint arXiv:2202.11317 (2022).

[33] A. Koudounas, M. La Quatra, L. Vaiani, L. Colomba, G. Attanasio, E. Pastor, L. Cagliero, E. Baralis, ITALIC: An Italian Intent Classification Dataset, in: Proc. INTERSPEECH 2023, 2023, pp. 2153–2157. doi:10.21437/Interspeech.2023-1980.