Divergence-aware Approaches to Mitigate Subgroup Disparities in Speech Models

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

[1]Politecnico di Torino, Italy

Abstract

Speech models often struggle with performance inconsistencies across different subgroups, leading to degraded accuracy for certain speaker demographics, accents, or recording conditions. These discrepancies may originate from multiple reasons, such as imbalanced training data, suboptimal representation learning, and limitations in model generalization. Addressing these issues allows for improving model robustness and reliability in real-world applications. We propose to mitigate performance disparities of subgroups that underperform, i.e., exhibit a *divergence*, relative to overall model performance. We tackle the performance disparities both via in-processing solutions, i.e., implementing mitigation measures during model development, and a post-processing one, refining already trained models. As in-processing solutions, we propose three approaches: divergence-aware regularization, targeted data augmentation, and contrastive learning (CLUES). Each method improves model learning in different ways: divergence-aware regularization adjusts training to focus on low-performing subgroups, targeted data augmentation generates synthetic variations to enhance model robustness, while CLUES refines latent representations. The post-processing strategy introduces a divergence-aware data acquisition method to prioritize acquiring real-world samples from underperforming subgroups.

Experiments on two spoken language understanding datasets and two languages demonstrate the effectiveness of our approaches in reducing subgroup disparities and improving overall performance, with CLUES and divergence-aware regularization leading to the highest improvement. The results demonstrate that integrating performance-aware training and representation learning strategies helps reducing inconsistencies across subgroups, making models more adaptable to diverse speech conditions.

Keywords

bias mitigation, spoken language understanding, speech processing, data acquisition, divergence

1. Introduction

Speech models are widely used in modern applications, including virtual assistants, transcription services, and accessibility tools [1, 2, 3, 4]. These models must handle a wide range of speech variations, including different accents, speaking styles, and recording conditions [5, 6, 7]. Despite their advancements, these models often exhibit performance disparities across different population subgroups. Studies have shown that factors such as gender, accent, speaking rate, and recording conditions can significantly impact the accuracy of these systems ([8, 9, 10, 11, 12, 13, 14, 15, 16]). These inconsistencies reduce the reliability of speech models and limit their

SEBD 2025: 33rd Symposium on Advanced Database System, June 16-19, 2025 - Ischia, Italy

*Corresponding author.

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

0000-0003-4386-0409 (A. Koudounas); 0000-0002-3664-4137 (E. Pastor); 0000-0001-8806-7979 (F. Giobergia); 0000-0001-9231-467X (E. Baralis)

ability to perform well across diverse real-world conditions. Several factors may contribute to these disparities. Differences in data distribution can lead to imbalanced learning, where models become more accurate for certain types of speech while struggling with others. Inadequate representation learning fails to capture the full spectrum of speech variations. Models may also struggle to generalize when encountering speech characteristics underrepresented during training. Addressing these issues allows for improving speech model robustness and ensuring that they perform consistently across different conditions.

Various methods have been proposed to address these challenges. Many approaches rely on manually identifying specific speech characteristics that might cause performance issues [17]. Some approaches use data augmentation [18], generating synthetic variations to improve model robustness. Others explore domain adaptation [19, 20], fine-tuning models on datasets that better represent specific speech characteristics. Adversarial training has also been used to make models more invariant to certain variations in speech [18]. While these techniques have improved fairness and robustness, they may overlook unexpected subgroups that emerge only after model evaluation. Moreover, performance disparities often occur at the intersection of multiple speech characteristics, making it difficult to address all sources of inconsistency through predefined subgroup selection alone.

Recent research has explored automated subgroup identification, using clustering techniques to detect data patterns where models underperform [12]. While these data-driven approaches help identify performance gaps, they often lack interpretability and do not clearly describe the underlying problems. Consequently, they do not provide insights into the specific sources of performance inconsistencies nor guidance for data acquisition for model improvement.

Our paper presents a framework addressing these limitations through four complementary methods. We propose to mitigate the performance disparities within data subgroups that deviate significantly, i.e., exhibit a *divergence*, from the overall model performance. We propose both post-processing and in-processing approaches. For post-processing, i.e., improving already trained models [21], we propose a targeted data acquisition to collect new real-world samples to fine-tune a pre-trained model, mitigating its disparities [22]. In-processing involves the implementation of mitigation measures during the model development phase [21]. As in-processing, we propose three techniques [23, 24]: divergence-aware regularization, targeted data augmentation, and contrastive learning. Divergence-aware regularization modifies the model loss function to emphasize underperforming subgroups during training. Targeted data augmentation increases the representation of these subgroups by applying transformations to existing samples. Finally, contrastive learning refines the model's internal representations by grouping similar samples closer together in latent space.

To evaluate these methods, we conduct experiments on two spoken language understanding datasets: Fluent Speech Commands (FSC) in English [25] and ITALIC [26] in Italian. We fine-tune transformer-based speech models and measure their performance using overall accuracy, subgroup performance divergence, and latent space analysis. Our results provide insights into the effectiveness of each method in reducing bias and improving performance.

2. Problematic Subgroup Identification on Interpretable Metadata

Speech models often exhibit inconsistent performance across different speaker groups. To address this issue, it is necessary first to identify and analyze these subgroups systematically. A challenge in subgroup identification is ensuring that the subgroups are interpretable, meaning they provide clear insights into why performance disparities occur. For instance, "young men in noisy scenarios" is an interpretable subgroup, allowing both understanding and intervention. To achieve this identification, we leverage the techniques of [27, 28, 29] that define subgroups as interpretable combinations of metadata such as speaker demographics, recording conditions, and task characteristics. In the following, we outline the definition of interpretable metadata and then the automatic identification of subgroups.

Interpretable Metadata. Speech datasets typically include a variety of metadata attributes that can influence model performance. Demographic attributes such as gender, age, and native language are among the most common factors affecting recognition accuracy. Beyond demographics, speech characteristics such as speaking rate and silence duration also impact recognition performance [30]. Faster speech or heavily accented pronunciation may introduce additional challenges, especially if the training data lacks sufficient diversity. In addition to speaker characteristics, recording conditions also contribute to subgroup disparities. Factors such as background noise, microphone type, and reverberation levels can create variations in audio quality, affecting model predictions. A model trained primarily on clean audio data may struggle when encountering noisy environments, leading to disparate performance outcomes for speakers who record in less controlled conditions. Task-specific metadata, such as intent categories in spoken language understanding, also play a role in subgroup performance. Certain intents or command structures may be more frequently represented in training data, resulting in better recognition accuracy compared to less frequent or more complex intent formulations.

Automatic subgroup identification. To systematically extract underperforming subgroups, we adopt DivExplorer [31, 32, 33]. DivExplorer identifies underperforming and interpretable subgroups by analyzing metadata attributes and measuring performance *divergence*, which quantifies how much a subgroup's performance deviates from the overall model performance.

Specifically, let D denote the dataset and A the set of metadata attributes. An *item* is defined as an attribute-value pair. For example, *gender=female* or *speaking rate=high* are items. A subgroup corresponds to the subset of data instances that satisfy one or more such items, represented as an *itemset* I. Given a statistic f (e.g., accuracy or error rate), the divergence $\Delta_f(I)$ of a subgroup identified by the itemset I is defined as: $\Delta_f(I) = f(I) - f(D)$. It indicates that the subgroup underperforms significantly relative to the dataset overall. A high negative divergence value indicates that the subgroup is significantly underperforming compared to the dataset as a whole. To ensure statistical reliability, subgroup discovery is constrained by a minimum support threshold, which filters out small subgroups where performance estimates may be unreliable. The subgroups are extracted by augmenting frequent pattern mining techniques, such as FP-Growth or Apriori, over the defined interpretable metadata, to also compute the divergence during the extraction process. By identifying subgroups with significant divergence, this method provides a structured way to analyze and mitigate performance inconsistencies. The

identified subgroups inform post-processing targeted data acquisition (§3.1) and in-processing techniques via regularization (§3.2), data augmentation (§3.3), and contrastive learning (§3.4).

3. Bias Mitigation Methods

Bias in speech models arises when performance varies significantly across different subgroups, often due to imbalanced representation in training data. To mitigate these disparities, various techniques have been proposed, broadly categorized into post-processing methods, which refine a trained model, and in-processing methods, which modify the training process itself. Post-processing methods are useful when fairness issues emerge after deployment, as they adjust model predictions or incorporate new data without requiring full retraining. In-processing methods, on the other hand, introduce fairness-aware mechanisms directly into the learning process to ensure balanced performance from the outset.

This study covers 4 bias mitigation techniques: one is a post-processing method, 3 are in-processing. For post-processing, we use targeted data acquisition, which enhances fairness by collecting additional subgroup-specific data. In-processing approaches include divergence-aware regularization, which modifies the loss function to prioritize underperforming subgroups; targeted data augmentation, which increases subgroup diversity through synthetic transformations; and contrastive learning (CLUES) to refine latent representations for improved fairness.

3.1. Post-Processing: Targeted Data Acquisition

Targeted data acquisition is a post-processing approach that improves subgroup performance by supplementing the training set with additional real-world examples from underperforming subgroups. This method identifies performance disparities after model deployment and retrains the model with newly collected data.

The process begins by evaluating the trained model to identify subgroups with significantly lower accuracy compared to the overall dataset. These subgroups are interpretable, ensuring that their characteristics are clearly defined. By guaranteeing interpretability, we can perform targeted data acquisition to acquire new speech samples that better represent them. These additional samples are then integrated into the dataset, and the model undergoes additional fine-tuning to improve its ability to generalize across all subgroups.

One of the key advantages of targeted data acquisition is its reliance on real-world speech variations rather than artificial data. This ensures that the model learns from natural speech patterns, accents, and recording conditions that were previously underrepresented. However, this method requires significant resources for data collection, annotation, and model retraining. Despite these challenges, targeted data acquisition is particularly valuable in deployed systems, where performance problems across groups only become apparent after real-world use.

3.2. In-Processing: Divergence-Aware Regularization

Traditional model learning functions optimize for overall performance, often overlooking subgroup disparities. Divergence-aware regularization is an in-processing technique that directly modifies the training process to subgroup learning. This approach dynamically adjusts

the loss function to focus on underperforming subgroups, ensuring they receive increased attention during training. In this method, the model training continuously monitors performance across different subgroups. If a subgroup exhibits significantly lower accuracy compared to the overall dataset, its samples are assigned higher loss weights during training. By amplifying the contribution of these samples, the model is encouraged to learn representations that better capture subgroup-specific variations.

Divergence-aware regularization is an effective solution for bias mitigation without requiring additional data. Since it operates directly on the training loss, it improves subgroup performance without altering the dataset size or introducing synthetic transformations.

3.3. In-Processing: Targeted Data Augmentation

Targeted data augmentation is another in-processing method that improves subgroup performance by artificially incrementing the training data for underperforming subgroups. Instead of collecting new samples, this approach applies synthetic transformations to the existing data to increase subgroup representation. Several augmentation techniques are commonly used in speech processing, including time stretching, which alters the speed of speech, pitch shifting, which changes the speaker's tone, and noise injection, which simulates different recording environments. These transformations create diverse variations of the same speech sample, allowing the model to become more robust to variations in speaking style, accent, or background noise. In our context, once the underperforming subgroups are identified, we apply targeted data augmentation techniques to increase their presence in the training set.

One key advantage of this approach is its efficiency, as augmentation can be applied easily to existing samples. However, this method does not introduce truly new linguistic or demographic diversity—it only manipulates existing samples. Despite this limitation, it serves as a cost-effective way to improve model robustness for underperforming subgroups.

3.4. In-Processing: Contrastive Learning (CLUES)

Contrastive learning has gained attention as an effective technique for refining the latent space representations of deep learning models. The CLUES (Contrastive Learning framework for Underperforming Subgroups) method applies contrastive loss to guide the model in learning more structured and subgroup-aware representations. Unlike regularization or data augmentation, which focuses on altering training behavior, contrastive learning reshapes the model's internal feature space to better distinguish between subgroups.

CLUES operates at three levels of contrastive learning. At the task level, it ensures that samples belonging to the same class are grouped closely together while separating samples from different classes. At the subgroup level, it clusters samples from the same subgroup while pushing apart those from different subgroups. Finally, at the error level, it groups correctly classified samples separately from misclassified ones within each subgroup. By optimizing these three objectives, CLUES improves how the model encodes subgroup-specific information, leading to improved subgroup performance.

A key advantage of CLUES is that it improves model representations at the subgroup level without requiring additional data by restructuring the way data is represented. By explicitly

shaping the latent space, CLUES reduces overlap between subgroup distributions, preventing the model from learning biased or entangled representations. However, contrastive learning introduces additional computational complexity. Despite this, experimental results show that CLUES provides the most effective technique for mitigating bias and improving performance.

Summary. Post-processing and in-processing methods offer distinct strategies for addressing bias in speech models. Targeted data acquisition, as a post-processing method, enhances subgroup performance by incorporating real-world samples into model fine-tuning. In contrast, in-processing methods adjust the training process to achieve improvement at the subgroup level without external data collection. The selection of an appropriate bias mitigation method depends on the specific requirements of the application, including available data and computational constraints objectives. In the next section, we outline the experimental setup used to evaluate these methods and analyze their effect on subgroup and overall model performance.

4. Results and Analysis

This section presents the results of applying the four bias mitigation methods, analyzing their impact on overall model performance, subgroup fairness, and latent space representations.

4.1. Experimental setup

Dataset and models. We conduct experiments on two spoken language understanding datasets: Fluent Speech Commands (FSC) [25] in English and ITALIC [26] in Italian. These datasets contain labeled utterances categorized by intent. The data is split into training, validation, and test sets, ensuring that speakers do not overlap between splits. To test the scenario of data acquisition of unseen samples, we also tested a configuration in which we use part of the original train set for actual training and a part for the data acquisition, denoted as held-out. We fine-tune wav2vec 2.0 [34] for FSC and XLS-R [35] for ITALIC . For our subgroup extraction with DivExplorer, we explored all subgroups with a minimum frequency of 0.03, following [22]. For both post-processing data acquisition and in-processing data augmentation, the hyperparameter K defines the top-K most challenging subgroups to attention. We report the results for K=2. Complete results with sensitivity analysis, ablation studies, and evaluations on emotion recognition and automatic speech recognition tasks are available in [22, 23, 24].

Metrics. We evaluate accuracy and macro F1 score to measure overall performance. For subgroup performance, we evaluate the maximum subgroup divergence (Δ_{max}), average divergence for the top-10 ($\Delta_{avg\text{-}10}$) underperforming subgroups, and the average divergence in absolute terms ($|\Delta_{avg\text{-}all}|$). We also performed a latent space analysis using the Silhouette Score to assess how well the model distinguishes between subgroups when adopting CLUES.

Baselines. We compare our mitigation methods when using our automatic identification approach against a set of alternative baselines that aim to identify challenging samples for model improvement. The *random* baseline selects samples randomly, serving as a control to highlight the effectiveness of subgroup-based selection. The *clustering* baseline follows [12], where challenging subgroups are identified using K-means clustering applied to acoustic embeddings. The clusters with the lowest performance are then used to determine the most challenging samples. The *KNN* baseline employs a K-Nearest Neighbors approach, where an utterance is

Table 1
Mean and standard deviation of three runs on FSC and ITALIC. For post-processing, we train on a subset of the training data (*original - no held out*) and use the held-out dataset for acquisition. For post-processing, we consider all training data (*original*) and then apply mitigation. Best results for each dataset are in **bold**, second-best underlined; best results for each dataset and strategy in light yellow.

DS	Method	Mitigation Strategy	Accuracy	F1 Macro	Δ^{-}_{max}	$\Delta^{-}_{avg\text{-}10}$	$\lvert\Delta_{avg\text{-}all}\rvert$
FSC	original - *no held out*	-	91.58±0.08	86.34±0.13	-70.09±0.26	-70.09±0.26	1.06±0.07
	w/ random	acquisition	92.56±0.44	90.25±0.60	-52.20±2.57	-51.11±2.19	0.97±0.02
	w/ KNN	acquisition	92.07±0.17	89.92±0.11	-49.90±0.33	-49.85±0.29	0.96±0.03
	w/ clustering	acquisition	89.77±0.88	87.02±0.15	-47.37±0.42	-47.34±0.42	0.94±0.04
	w/ error-driven	acquisition	95.71±0.74	94.06±0.83	-48.13±0.39	-48.02±0.36	0.92±0.04
	ours - w/DivExplorer	acquisition	96.55±0.08	94.71±0.12	-40.60±0.35	-40.28±0.36	0.81±0.03
	original - *all*	-	93.42±0.17	93.11±0.17	-53.18±0.15	-50.89±0.09	0.37±0.01
	w/ random	target data++	94.91±0.87	94.46±0.86	-42.62±2.94	-42.51±2.88	0.36±0.24
	w/ KNN	target data++	96.72±0.34	96.15±0.39	-40.01±1.59	-39.59±1.57	0.31±0.08
	w/ clustering	target data++	97.85±0.37	97.59±0.65	-37.57±2.68	-37.21±2.49	0.24±0.11
	ours - w/DivExplorer	target data++	98.46±0.11	98.42±0.17	-27.51±0.56	-27.12±0.52	0.21±0.08
	w/ random	regularization	96.46±0.56	96.29±0.66	-41.31±7.00	-41.31±7.00	0.79±0.94
	w/ KNN	regularization	97.55±0.28	97.38±0.24	-38.29±2.34	-38.02±2.25	0.53±0.06
	w/ clustering	regularization	97.88±0.33	97.65±0.57	-36.95±8.44	-36.28±8.21	0.13±0.02
	ours - w/DivExplorer	regularization	98.47±0.11	98.43±0.14	-24.49±0.57	-24.49±0.57	0.11±0.01
	w/ clustering	CLUES	98.57±0.15	98.51±0.14	-21.41±0.39	-21.12±0.25	0.12±0.02
	ours - w/DivExplorer	CLUES	**98.79±0.10**	**98.76±0.10**	**-17.58±0.43**	**-17.44±0.38**	**0.05±0.04**
ITALIC	original - *no held out*	-	73.79±0.32	68.08±0.37	-47.63±1.93	-47.52±1.94	0.60±0.01
	w/ random	acquisition	75.32±0.63	70.72±0.58	-47.00±0.81	-46.86±0.80	0.48±0.02
	w/ KNN	acquisition	75.56±0.57	70.21±0.54	-46.11±0.93	-46.02±0.92	0.39±0.02
	w/ clustering	acquisition	74.05±0.33	69.09±0.75	-45.02±2.02	-44.91±2.01	0.37±0.08
	w/ error-driven	acquisition	77.14±0.52	72.65±0.63	-46.97±1.15	-46.84±1.07	0.45±0.04
	ours - w/DivExplorer	acquisition	77.40±0.24	72.51±0.14	-31.75±0.55	-31.71±0.55	0.34±0.03
	original - *all*	-	75.71±0.36	73.22±0.33	-47.54±0.79	-47.36±0.76	0.15±0.03
	w/ random	target data++	76.06±0.29	73.36±0.77	-45.82±1.89	-45.34±1.72	0.13±0.09
	w/ KNN	target data++	77.15±0.21	74.03±0.24	-37.87±0.89	-37.12±0.83	0.12±0.04
	w/ clustering	target data++	77.81±0.56	74.19±0.49	-36.73±2.53	-36.19±2.27	0.08±0.02
	ours - w/DivExplorer	target data++	78.01±0.49	74.74±0.35	-30.49±1.77	-30.02±1.52	0.05±0.03
	w/ random	regularization	77.47±0.22	72.76±0.22	-45.11±1.41	-44.99±1.40	0.10±0.01
	w/ KNN	regularization	77.96±0.19	74.12±0.23	-36.39±1.17	-36.14±1.09	0.07±0.02
	w/ clustering	regularization	78.01±0.45	74.45±0.35	-32.81±2.35	-32.73±2.32	0.05±0.03
	ours - w/DivExplorer	regularization	78.07±0.53	74.85±0.30	**-30.10±1.71**	**-29.64±1.70**	**0.01±0.04**
	w/ clustering	CLUES	**80.56±0.55**	76.10±0.32	-43.01±0.89	-42.79±0.86	0.03±0.02
	ours - w/DivExplorer	CLUES	79.23±0.81	**76.72±0.20**	-40.15±0.96	-40.01±0.97	**0.01±0.02**

considered challenging if its nearest neighbors in the validation set are frequently misclassified, with K optimized per dataset. Finally, the *error-driven* baseline, close to [36], selects misclassified instances from the held-out set and incorporates them into training. We evaluate this baseline only for post-processing since training loss inherently accounts for errors during learning.

4.2. Experimental results

Overall Performance. We report the results in Table 1. The four proposed bias mitigation methods lead to varying degrees of improvement in model accuracy and fairness. Divergence-aware regularization and contrastive learning (CLUES) achieve highest overall accuracy while allowing the highest reductions in subgroup disparities. Coupling any strategy with our

identification methodology generally always achieves the best results (light yellow).

On the FSC dataset, when using the full training set (*original-all*), the baseline wav2vec 2.0 model achieves an accuracy of 93.42% and an F1 macro score of 93.11%, but exhibits high divergence across subgroups. After applying mitigation strategies, CLUES improves overall accuracy to 98.79% and reduces subgroup divergence significantly. Divergence-aware regularization similarly enhances subgroup performance while maintaining a competitive accuracy of 98.5%, while targeted data augmentation yields more moderate improvements, particularly benefiting subgroups with lower representation. On ITALIC, the baseline XLS-R model achieves 73.22% F1 Macro (*original - all*). Divergence-aware regularization and contrastive learning improve overall performance to 74.85% for the former and to 76.10% and 76.72% when using CLUES coupled with clustering or our identification approach based on DivExplorer.

Subgroup Performance. We assess how well the methods reduce subgroup disparities. Before mitigation, the baseline FSC model on overall data has a Δ_{max} of 53.18% (i.e., the least accurate subgroup performs significantly worse than the global accuracy of 93.42%). CLUES reduces the most this divergence, down to 17.58%. Divergence-aware regularization also substantially reduces Δ_{max} to 24.49%, confirming its effectiveness in addressing subgroup imbalances. For ITALIC, the baseline XLS-R model on overall data starts with a Δ_{max} of 47.54%. Contrastive learning and divergence-aware regularization reduce this gap to 40.15% and 30.10%.

Latent Space Analysis. We use the Silhouette Score to investigate the impact of bias mitigation on the latent space representations. A higher Silhouette Score indicates that the model better separates subgroups, reflecting improved internal representations of speech variations.

The baseline FSC model achieves a Silhouette Score of 0.737. CLUES improves this to 0.894, demonstrating that targeting subgroup representation learning significantly enhances the model's ability to distinguish between subgroups. A similar pattern is observed in the ITALIC dataset, where contrastive learning improves Silhouette Scores from 0.319 to 0.539. This suggests that models trained with contrastive objectives learn more structured and subgroup-aware representations, contributing to improvements in subgroup performance. A complete analysis of model representations can be found in [24].

5. Conclusions

This paper outlined a framework for improving speech model performance by identifying and mitigating subgroup disparities, leveraging interpretable metadata to systematically detect underperforming, i.e., *divergent*, subgroups. We explored four mitigation techniques: the *post-processing* targeted data acquisition and the *in-processing* divergence-aware regularization, targeted data augmentation, and contrastive learning (CLUES). Each method addressed performance inconsistencies differently, either by enhancing model training, refining latent representations, or incorporating subgroup-specific data. The experimental results show that CLUES and the divergence-aware regularization are the most effective in reducing subgroup disparities. Moreover, CLUES enhances latent space representations. The findings highlight the value of adopting divergence-aware subgroup identification in speech model development.

Acknowledgments

This work is partially supported by the FAIR - Future Artificial Intelligence Research and received funding from the European Union NextGenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013) and the spoke "FutureHPC & BigData" of the ICSC - Centro Nazionale di Ricerca in High-Performance Computing, Big Data and Quantum Computing funded by the European Union - NextGenerationEU. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Declaration on Generative AI

During the preparation of this work, the author(s) used, Grammarly, ChatGPT to check grammar and spelling, paraphrase and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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.

[2] M. La Quatra, A. Koudounas, E. Baralis, S. M. Siniscalchi, Speech analysis of language varieties in italy, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 15147–15159.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: International Conference on Machine Learning, PMLR, 2023, pp. 28492–28518.

[4] A. Koudounas, G. Ciravegna, M. Fantini, E. Crosetti, G. Succo, T. Cerquitelli, E. Baralis, et al., Voice disorder analysis: a transformer-based approach, in: INTERSPEECH, ISCA, 2024, pp. 3040–3044.

[5] L. Vaiani, A. Koudounas, M. La Quatra, L. Cagliero, P. Garza, E. Baralis, Transformer-based non-verbal emotion recognition: Exploring model portability across speakers' genders, in: Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge, MuSe' 22, Association for Computing Machinery, New York, NY, USA, 2022, p. 89–94. URL: https://doi.org/10.1145/3551876.3554801. doi:10.1145/3551876.3554801.

[6] T. Feng, S. Narayanan, Foundation model assisted automatic speech emotion recognition: Transcribing, annotating, and augmenting, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2024, pp. 12116–12120.

[7] A. Koudounas, M. La Quatra, S. M. Siniscalchi, E. Baralis, voc2vec: A foundation model for

non-verbal vocalization, in: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, pp. 1–5. doi:10.1109/ICASSP49660.2025.10890672.

[8] J. P. Bajorek, Voice recognition still has significant race and gender biases, Harvard Business Review 10 (2019).

[9] 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 (2020).

[10] Z. Mengesha, C. Heldreth, M. Lahav, J. Sublewski, E. Tuennerman, "i don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on african americans, Frontiers in Artificial Intelligence (2021) 169.

[11] 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.

[12] 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.

[13] Z. Liu, I.-E. Veliche, F. Peng, Model-based approach for measuring the fairness in asr, in: ICASSP, IEEE, 2022.

[14] A. Koudounas, E. Pastor, V. Mazzia, M. Giollo, T. Gueudre, E. Reale, G. Attanasio, L. Cagliero, S. Cumani, L. De Alfaro, E. Baralis, D. Amberti, Leveraging confidence models for identifying challenging data subgroups in speech models, in: 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), 2024, pp. 134–138. doi:10.1109/ICASSPW62465.2024.10626001.

[15] S. Feng, B. M. Halpern, O. Kudina, O. Scharenborg, Towards inclusive automatic speech recognition, Computer Speech & Language 84 (2024) 101567.

[16] A. Koudounas, F. Giobergia, Houston we have a divergence: A subgroup performance analysis of asr models, in: 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), 2024, pp. 812–813. doi:10.1109/ICASSPW62465.2024.10626156.

[17] O. Niebuhr, A. Michaud, Speech data acquisition: the underestimated challenge, KALIPHO-Kieler Arbeiten zur Linguistik und Phonetik 3 (2015) 1–42.

[18] Y. Zhang, Y. Zhang, B. M. Halpern, T. Patel, O. Scharenborg, Mitigating bias against non-native accents, in: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, volume 2022, 2022, pp. 3168–3172.

[19] 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.

[20] H. Shen, Y. Yang, G. Sun, R. Langman, E. Han, J. Droppo, A. Stolcke, Improving fairness in speaker verification via group-adapted fusion network, in: ICASSP, IEEE, 2022.

[21] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM computing surveys (CSUR) 54 (2021) 1–35.

[22] A. Koudounas, E. Pastor, G. Attanasio, Luca, 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. 1–5. doi:10.1109/ICASSP48485.2024.10446326.

[23] A. Koudounas, E. Pastor, L. de Alfaro, E. Baralis, Mitigating subgroup disparities in speech models: A divergence-aware dual strategy, IEEE Transactions on Audio, Speech and Language Processing 33 (2025) 883–895. doi:10.1109/TASLPRO.2025.3539429.

[24] A. Koudounas, F. Giobergia, E. Pastor, E. Baralis, A contrastive learning approach to mitigate bias in speech models, in: Interspeech 2024, 2024, pp. 827–831. doi:10.21437/Interspeech.2024-1219.

[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] 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.

[27] 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.

[28] 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.

[29] A. Koudounas, E. Pastor, E. Baralis, Assessing speech model performance: A subgroup perspective, in: SEBD 2024: 32nd Symposium on Advanced Database System, volume 3741, CEUR Workshop Proceedings, 2024, pp. 101–111. URL: https://ceur-ws.org/Vol-3741/paper64.pdf.

[30] E. Pastor, A. Koudounas, G. Attanasio, D. Hovy, E. Baralis, Explaining speech classification models via word-level audio segments and paralinguistic features, in: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 2221–2238. URL: https://aclanthology.org/2024.eacl-long.136/.

[31] 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.

[32] E. Pastor, A. Gavgavian, E. Baralis, L. de Alfaro, How divergent is your data?, Proc. VLDB Endow. 14 (2021) 2835–2838. URL: https://doi.org/10.14778/3476311.3476357. doi:10.14778/3476311.3476357.

[33] E. Pastor, E. Baralis, L. de Alfaro, A hierarchical approach to anomalous subgroup discovery, in: 2023 IEEE 39th international conference on data engineering (ICDE), IEEE, 2023, pp. 2647–2659. doi:10.1109/ICDE55515.2023.00203.

[34] 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. URL: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

[35] A. Babu, et al., XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale, in: Proc. Interspeech 2022, 2022. doi:10.21437/Interspeech.2022-143.

[36] R. Magar, A. B. Farimani, Learning from mistakes: Sampling strategies to efficiently train machine learning models for material property prediction, Computational Materials Science 224 (2023).