Privacy Preserving Data Selection for Bias Mitigation in Speech Models

Privacy Preserving Data Selection for Bias Mitigation in Speech Models

Alkis Koudounas[†] **Eliana Pastor**[†] **Vittorio Mazzia**[‡] **Manuel Giollo**[‡]
Thomas Gueudre[‡] **Elisa Reale**[‡] **Luca Cagliero**[†] **Sandro Cumani**[†]
Luca de Alfaro[⋆] **Elena Baralis**[†] **Daniele Amberti**[‡]

[†]Politecnico di Torino, Italy [‡]Amazon AGI, Italy [⋆]University of California, Santa Cruz
alkis.koudounas@polito.it

Abstract

Effectively selecting data from population subgroups where a model performs poorly is crucial for improving its performance. Traditional methods for identifying these subgroups often rely on sensitive information, raising privacy issues. Additionally, gathering such information at runtime might be impractical. This paper introduces a cost-effective strategy that addresses these concerns. We identify underperforming subgroups and train a model to predict if an utterance belongs to these subgroups without needing sensitive information. This model helps mitigate bias by selecting and adding new data, which is labeled as challenging, for re-training the speech model. Experimental results on intent classification and automatic speech recognition tasks show the effectiveness of our approach in reducing biases and enhancing performance, with improvements in reducing error rates of up to 39% for FSC, 16% for ITALIC, and 22% for LibriSpeech.

1 Introduction

Speech models, such as those deployed in Automatic Speech Recognition (ASR) and Intent Classification (IC), often face challenges leading to subpar performance within specific population subgroups, as shown by recent studies (Dheram et al., 2022; Koudounas et al., 2023b; Liu et al., 2022). Identifying and addressing these subgroups is crucial for improving model robustness and ensuring fairness across diverse populations (Zhang et al., 2022; Shen et al., 2022; Koudounas et al., 2024a, 2025).

However, traditional methods for subgroup identification, which rely on demographic attributes like age, gender, and accent, raise privacy concerns since collecting such sensitive information during testing or deployment is often impractical or undesirable (Zhang et al., 2022; Padmanabhan et al., 1996). Recently, significant efforts have focused on enhancing the protection of user data, especially in relation to voice (Tran and Soleymani, 2023; Chen et al., 2024; Hashimoto et al., 2016; Panariello et al., 2024). While newer approaches have introduced speaker embeddings to tackle this issue (Dheram et al., 2022; Veliche and Fung, 2023), they continue to struggle, especially regarding their interpretability.

To address these challenges and reduce the dependence on sensitive demographic data, we propose the use of a Challenging Subgroup Identification (CSI) model, as introduced in Koudounas et al. (2024d), which is built on top of a Confidence Model (CM). Confidence scores, derived either from model-specific uncertainty estimates or through auxiliary CMs trained to predict error rates (Abdar et al., 2021; Swarup et al., 2019), are crucial in evaluating model reliability. Integrating CMs has been proven to help close performance gaps among demographic cohorts (Dheram et al., 2022). The CSI model identifies difficult subgroups without relying on demographic information, thus improving interpretability and transparency. We first apply automatic identification methods (Koudounas et al., 2024c) to detect challenging human-understandable subgroups and then fine-tune the CSI to predict these subgroups based on the confidence model outputs. This allows the CSI to identify performance challenges without compromising user privacy, enabling fair and responsible deployment of speech models.

We propose utilizing the CSI to mitigate model disparities in data subgroups by selecting additional labeled data tailored to these cohorts. Subset selection of data in speech processing serves various purposes, including (i) budget-constrained sampling (Lin and Bilmes, 2009; Wei et al., 2014a,b; Park et al., 2022), (ii) human annotation, especially relevant for new languages or dialects where audio has not been transcribed yet (Hakkani-Tür et al., 2002; Lamel et al., 2002; Kemp and Waibel, 1998), and (iii) bias mitigation in speech models (Dheram

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 738–748
July 28-30, 2025

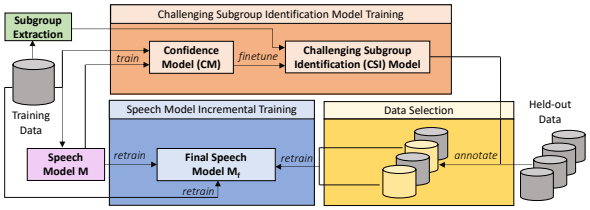

Figure 1: **Schema of the proposed pipeline.** We train the CSI model by fine-tuning a CM to predict the challenging subgroup an utterance belongs to (Koudounas et al., 2024d). We augment the original train set with the utterances of the held-out set labeled as challenging by the CSI to incrementally train the speech model.

et al., 2022; Koudounas et al., 2024b).

We focus on using the CSI to address subgroup disparities by selecting data specific to challenging subgroups. Few recent works have explored the data selection and acquisition of automatically identified challenging groups. The authors of Dheram et al. (2022) first derive challenging clusters of embedding representations and acquire data accordingly, while Koudounas et al. (2024b) considers interpretable subgroups defined over metadata (e.g., gender, age, speaking rate of the utterances). Their work shows the benefit of interpretable subgroups over not interpretable clusters in mitigating subgroup disparities and improving performance. However, the approach requires knowing sensitive information for the data to be acquired. In contrast, our approach offers interpretability without the need for sensitive data. This privacy-preserving methodology ensures fairness while maintaining transparency and improving model performance.

Experimental results on FSC (Lugosch et al., 2019) and ITALIC (Koudounas et al., 2023a) datasets for IC, and on LibriSpeech (Panayotov et al., 2015) for ASR, validate our methodology. Our approach obtains a reduction in Intent Error Rate (IER) up to 39% for FSC and 16% for ITALIC and a 22% decrease in Word Error Rate (WER) for LibriSpeech. We observe lower error rates and higher macro F1 scores compared to various baselines employing KNN, clustering (Dheram et al., 2022), and model mistakes (Magar and Farimani, 2023) to guide the data selection process. By avoiding demographic data collection, we offer a privacy-aware alternative that enhances both fairness and model performance, thus remaining competitive with data selection strategies that traditionally rely on sensitive information (Koudounas et al., 2024b).

This work addresses a critical challenge for commercial speech recognition systems, which must balance performance improvements with increasing privacy concerns and regulations. Our approach enables organizations to deploy fairer speech models in production environments without requiring the collection of sensitive user data, thus aligning with real-world deployment constraints across various industries. The main contributions of this work are threefold: (i) we propose a novel privacy-preserving approach to enhance overall model performance and mitigate subgroup disparities without the need to access or collect sensitive information; (ii) we address both the drawbacks of current mitigation approaches that rely on the availability of metadata, demographic included, at deployment time or on acoustic embedding clustering, which results in non-interpretable groups; and (iii) we demonstrate the effectiveness of our solution on two speech tasks, three datasets, two languages, and a wide range of existing baseline approaches.

2 Methodology

We consider a speech model $\mathcal{M}$ designed for tasks such as IC or ASR. We aim to improve its performance by mitigating biases observed in population subgroups. Our approach consists of two main steps, as shown in Figure 1. We first train a Challenging Subgroup Identification (CSI) model that predicts if an utterance belongs to a *challenging* subgroup for model $\mathcal{M}$. We then re-train the speech model $\mathcal{M}$ by acquiring new data that the CSI model predicted to be challenging. The proposed framework is designed with practical deployment considerations in mind, requiring minimal additional computational overhead while enabling continuous improvement of production systems. By focusing on challenging subgroups rather than individual errors, our approach allows for more efficient model updates in real-world applications.

Challenging Subgroup Identification model. The CSI model was introduced in Koudounas et al. (2024d); we summarize here its main characteristics. It predicts whether an utterance is challenging for a model and, if so, identifies the challenging subgroup it belongs to. The model consists of two components: a pre-trained confidence model (CM) and ground-truth challenging subgroups.

Confidence model. Given an input dataset $\mathcal{X}$, we define a transformed dataset $\mathcal{Z}$ for training the CM. This dataset consists of input features and error-based target labels. Such features include (i)

uncertainty measures, e.g., n-best list length and output probabilities, (ii) acoustic embeddings from the model's hidden states, and (iii) speech metadata like word count, pauses, and speaking rate. Each utterance is labeled 1 if $\mathcal{M}$ predicts it correctly and 0 otherwise. In ASR, the label 1 corresponds to a perfect WER of 0.0. We train the CM on $\mathcal{Z}$ by splitting it into standard training, validation, and test subsets.

Challenging subgroup. We then identify challenging subgroups from the dataset using the DivExplorer (Pastor et al., 2021) method as described in Koudounas et al. (2023b). DivExplorer analyzes interpretable metadata describing utterances to extract all *frequent* subgroups and calculate their *divergence*, i.e., difference, in performance from the overall dataset. Subgroups are defined as "frequent" based on a set support threshold. First, we enrich the dataset with metadata, including demographic, speaking or recording conditions, and task-specific information, which is assumed to be available during training. This metadata allows us to develop a model that accounts for sensitive attributes, which may be unavailable at runtime. Each subgroup is defined by metadata-value pairs (e.g., {*gender=female, duration>10s*}). We focus on the top K challenging subgroups with below-average performance compared to overall behavior.

CSI model. We finally train the CSI model to predict the challenging subgroup for each utterance by fine-tuning the CM. The transformed dataset $\mathcal{Z}$ is labeled with the IDs of challenging subgroups. Specifically, each utterance in $\mathcal{Z}$ is annotated with (i) the ID of its most divergent challenging subgroup or (ii) a non-challenging ID if it does not belong to any challenging subgroup. Unlike Koudounas et al. (2024d), which used a multi-class setting to predict K distinct subgroups, we collapse the K challenging subgroups into a unique class, as our goal is to use CSI to acquire new data that challenges the model.

Bias Mitigation. We aim to enhance the performance of model $\mathcal{M}$, both overall and within specific data subgroups. Rather than indiscriminately acquiring and retraining on new data, a recent study highlighted the effectiveness of a more targeted approach to data acquisition (Koudounas et al., 2024b). Building on this paradigm, we use the CSI to guide the acquisition process, specifically targeting utterances without the need for sensitive information such as demographic data. This privacy-preserving method enables subgroup-based, focused data selection, allowing us to acquire new data in a way that directly addresses model disparities while safeguarding user privacy.

We start with a set of held-out utterances not used in training models $\mathcal{M}$, CM, and CSI. These utterances are labeled with the CSI model to determine if they likely belong to a challenging subgroup. We enhance the training data by including those identified as challenging and re-train model $\mathcal{M}$ by fine-tuning it on the initial training dataset combined with the selected data (referred to as model $\mathcal{M}_f$ in Figure 1).

3 Experimental Setup

This section details datasets, models, metrics, training procedures, and baselines used for the experiments[1]. Further details can be found in Appendix A and in the project repository.

Datasets. We assess our approach on three datasets: Fluent Speech Commands (FSC) (Lugosch et al., 2019) for English and ITALIC (Koudounas et al., 2023a) for Italian for the IC task, and LibriSpeech (Panayotov et al., 2015) for ASR. More details on the datasets and the available and extracted metadata are in Appendix A.1.

Confidence model. Following Koudounas et al. (2024d), the CM architecture features two hidden layers with GELU activation functions, dropout, and normalization layers, initialized using the Kaiming normal technique. The training details can be found in Appendix A.2.

Models and training procedure. We consider two transformer-based speech models for IC, wav2vec 2.0 (Baevski et al., 2020) base for FSC and XLS-R (Babu and et al., 2022) for ITALIC, and Whisper (Radford et al., 2023) base for LibriSpeech. Each IC model undergoes fine-tuning by adding a final classification layer to the encoder architecture. For ASR, the entire Whisper model is fine-tuned. More details on models, training hyper-parameters, and hardware used are given in Appendix A.3.

We partition our datasets into training, held-out, validation, and test sets. The validation and test sets remain consistent with the original dataset splits, while the training set is divided into 80% for training and 20% held out. We use the training set for model training and the validation set to identify challenging subgroups. We also train and validate

[1]Code: github.com/koudounasalkis/CSI-MIT

the CM and CSI models on these partitions. Subsequently, data samples are acquired using stratified sampling from the held-out set to retrain the model. We evaluate the overall and subgroup model performance on the test set. While using additional external data would be a practical and optimal choice for improving the model, for experimental purposes, the 20% held-out data is adequate to demonstrate our approach's effectiveness. It also serves as a good proxy for the overall data distribution, allowing us to assess the CSI's performance.

To ensure a fair comparison, we consider each approach separately and determine the number of N possible samples to acquire from the held-out set. Apart from the random baseline, all other baselines may limit the number of data identified as challenging due to the limited size of the held-out set. We then identify the minimum value of N across all methods and select this consistent number of samples for all approaches. This approach disentangles the impact of the number of added instances from the method itself. As a result, any improvement in the final performance can be attributed to the specific selection method rather than the number of added instances.

Metrics. We assess model performance using Intent Error Rate (IER) and F1 Macro scores for IC and WER for ASR. We also evaluate performance at the subgroup level, considering the IER and WER for the top-K challenging subgroups, with K in the range $[2, 5]$.

Baselines. We benchmark our approach against six baselines.

Random baseline. We randomly add instances from the held-out dataset to the training data.

KNN baseline. We employ a K-Nearest Neighbors classifier. We identify the K closest utterances, via standard Euclidean distance, from the training set for each instance in the held-out set, represented in the same input space as in our methodology. The selection of K is based on maximizing the performance, i.e., identifying challenging subgroups on the validation set. We determine if an utterance is challenging or not through majority voting among these neighbors. Predicted challenging instances are included in the retraining process.

Cluster-based baseline. We adopt an unsupervised clustering approach inspired by Dheram et al. (2022) to identify challenging subgroups. First, we extract acoustic embeddings from audio samples using the last layer of the Whisper model, with a fixed length for each utterance. We then apply K-means clustering with standard settings to group these embeddings into similar clusters. Consistent with Dheram et al. (2022), we use 50 clusters, as this number has been shown to adequately capture speech characteristics pertinent to ASR. Finally, we select the clusters with the poorest performance for targeted data acquisition.

CM-based baseline. We use the CM to label the utterances and include samples labeled as erroneous in the training data.

We further employ two baselines that work as *oracles*, as they assume the knowledge of ground truth labels or metadata, demographics included.

Supervised oracle (`S-Oracle`). Similarly to the methodology proposed in Magar and Farimani (2023), we use an erroneous-sample-driven approach that incorporates instances predicted erroneously by the model into the augmented training data. This baseline assumes the prior knowledge of the ground truth labels for the tasks, hence serving as the oracle for the CM-based baseline.

Metadata-based oracle (`M-Oracle`). We adopt the approach described in Koudounas et al. (2024b), which assumes access to metadata, including sensitive demographic information, for the samples in the held-out set to be acquired. This approach represents the oracle for our proposal since, in our work, we use the CSI to predict the challenging subgroups without accessing such metadata.

4 Results and Discussion

We evaluate the performance of our targeted data selection approach on three datasets and two tasks: FSC and ITALIC for the IC task and LibriSpeech for ASR. Table 1 focuses on the results on FSC. Our method effectively addresses performance disparities by reducing the IER of the top-K subgroups of about 50% for $K = 2$ and more than 60% for $K = 5$ w.r.t. the original fine-tuned model. This mitigation, in turn, leads to overall performance enhancement, with a 39% reduction in IER and almost 10% improvement in F1 macro scores. These results outperform all the considered baselines for every number K of subgroups considered.

We also test our approach against the two oracles, which use demographic-sensitive metadata and ground truth labels. Our methodology serves as a reliable proxy when compared to the metadata-based oracle (`M-Oracle` in Table 1). Even without demographic information, our method consistently

Table 1: **FSC, wav2vec 2.0 base.** Mean ±std of three runs. `K` indicates the number of challenging subgroups considered, `N` is the number of samples selected. We compare the results of the `Original` fine-tuning procedure, the baselines, our `CSI`, and the two oracles (`M-Oracle` considering metadata, `S-Oracle` leveraging supervised labels). Best results for each number of subgroups K are highlighted with light-blue. Best results with oracles in **bold**.

K	N	Approach	IER (%) ↓	F1 Macro (%) ↑	IER top-K (%) ↓
-	18506	Original	8.42±0.08	86.34±0.13	67.63±0.08 (K = 2)
2	+223	Random	9.19±0.03	88.48±0.05	65.90±0.22
		KNN	7.93±0.07	89.92±0.10	59.90±0.23
		Clustering (Dheram et al., 2022)	7.06±0.07	91.82±0.15	47.35±0.42
		CM	6.87±0.04	93.93±0.05	52.24±0.35
		CSI (*ours*)	5.17±0.03	**94.87±0.03**	34.04±0.21
		S-Oracle (Magar and Farimani, 2023)	5.29±0.02	94.06±0.04	47.47±0.39
		M-Oracle (Koudounas et al., 2024b)	**4.46±0.08**	94.81±0.09	**32.95±0.36**
-	+4606	All data	6.58±0.17	93.11±0.17	55.11±0.24 (K = 2)
3	+361	Random	9.41±0.05	88.15±0.09	49.44±0.38
		KNN	8.25±0.09	89.12±0.14	39.30±0.36
		Clustering (Dheram et al., 2022)	7.19±0.06	91.06±0.09	37.15±0.39
		CM	6.15±0.05	92.30±0.07	38.80±0.43
		CSI (*ours*)	5.25±0.04	94.21±0.07	23.17±0.23
		S-Oracle (Magar and Farimani, 2023)	5.60±0.04	93.43±0.04	51.17±0.35
		M-Oracle (Koudounas et al., 2024b)	**5.12±0.04**	**94.41±0.06**	**22.89±0.12**
4	+397	Random	9.45±0.11	88.09±0.10	36.44±0.27
		KNN	8.29±0.02	89.51±0.07	25.50±0.29
		Clustering (Dheram et al., 2022)	7.42±0.07	90.89±0.08	36.08±0.31
		CM	6.59±0.04	91.75±0.05	38.19±0.25
		CSI (*ours*)	5.31±0.03	94.19±0.05	19.89±0.21
		S-Oracle (Magar and Farimani, 2023)	5.84±0.06	93.44±0.06	46.40±0.33
		M-Oracle (Koudounas et al., 2024b)	**5.19±0.06**	**94.25±0.07**	**18.72±0.17**
5	+467	Random	9.58±0.10	88.04±0.10	34.80±0.39
		KNN	8.31±0.03	89.50±0.06	21.24±0.23
		Clustering (Dheram et al., 2022)	7.68±0.06	90.61±0.05	29.75±0.27
		CM	6.70±0.05	91.69±0.03	25.34±0.23
		CSI (*ours*)	5.39±0.06	94.05±0.04	14.55±0.08
		S-Oracle (Magar and Farimani, 2023)	5.85±0.06	**94.76±0.03**	46.94±0.25
		M-Oracle (Koudounas et al., 2024b)	**5.28±0.04**	94.08±0.06	**14.01±0.11**
-	+4606	All data	6.58±0.17	93.11±0.17	39.78±0.12 (K = 5)

yields comparable results across different K values. Notably, the top-K most challenging subgroups often involve sensitive attributes, e.g., age and gender. For FSC, in the top-2 we find the subgroup of male speakers aged 41-65 who speak quickly. Further examples of retrieved subgroup composition can be found in Appendix B. This demonstrates our approach's effectiveness in identifying challenging subgroups and acquiring data accordingly, all while avoiding direct access to sensitive information.

The supervised oracle (`S-Oracle`), which relies on ground truth labels, serves as a reference for the CM-based strategy. This oracle and our CSI achieve comparable overall intent error rates and F1 macro score, with our approach performing slightly better and showing improved IER for the top-K subgroups (IER top-K). We attribute this performance enhancement to our model's awareness of disparities within distinct population subgroups, which enables targeted retraining. Conversely, the supervised oracle disregards the information about the challenging subgroups, focusing on the samples that the model will predict incorrectly.

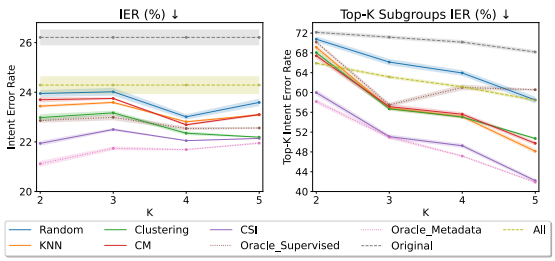

Figure 2: **ITALIC, `XLS-R` large.** Intent Error Rate (IER) and Top-K Subgroups IER for $K \in [2, 5]$.

Similar considerations also apply to ITALIC and

Table 2: **LibriSpeech, Whisper base**. Mean ±std of three runs. Best results for each number of subgroups K in light-blue, best results w/ oracles in **bold**.

K	N	Approach	WER ↓	WER top-K ↓
-	83211	Original	8.05±0.05	25.91±0.98 (K = 2)
2	+6912	Random	7.96±0.29	25.02±0.44
		KNN	7.80±0.04	18.44±0.32
		Clustering	7.33±0.08	14.05±0.38
		CM	7.70±0.09	14.86±0.27
		CSI (*ours*)	7.25±0.06	**12.33±0.15**
		S-Oracle	7.28±0.09	24.17±0.29
		M-Oracle	**7.22±0.06**	12.51 ±0.09
-	+20803	All data	6.31±0.07	17.46±0.87 (K = 2)
3	+8120	Random	7.71±0.31	22.15±0.41
		KNN	7.55±0.05	16.29±0.28
		Clustering	7.08±0.10	13.09±0.31
		CM	7.49±0.07	13.01±0.23
		CSI (*ours*)	6.81±0.08	10.97±0.17
		S-Oracle	6.87±0.07	21.86±0.32
		M-Oracle	**6.80±0.05**	**10.94±0.11**
4	+9958	Random	7.40±0.24	20.43±0.33
		KNN	7.33±0.04	14.84±0.19
		Clustering	6.81±0.08	12.55±0.24
		CM	7.21±0.05	12.56±0.18
		CSI (*ours*)	6.48±0.07	10.16±0.15
		S-Oracle	6.47±0.09	19.74±0.29
		M-Oracle	**6.43±0.05**	**10.15±0.09**
5	+12026	Random	7.14±0.09	17.52±0.31
		KNN	7.03±0.04	12.77±0.16
		Clustering	6.42±0.07	11.19±0.26
		CM	6.81±0.05	11.04±0.19
		CSI (*ours*)	6.32±0.04	9.33±0.13
		S-Oracle	6.34±0.05	15.01±0.26
		M-Oracle	**6.31±0.04**	**9.32±0.08**
-	+20803	All data	6.31±0.07	12.24±0.79 (K = 5)

LibriSpeech. Figure 2 visually illustrates the intent error rates both at the overall (IER) and subgroup (Top-K Subgroups IER) levels for the ITALIC dataset. The error rates are higher w.r.t. FSC, as the Italian dataset is more complex, and the multilingual XLS-R model achieves *per se* worst initial scores. Nonetheless, our approach consistently outperforms baselines and the supervised oracle while exhibiting comparable results to the metadata-based one. These findings emphasize the robustness and effectiveness of the proposed methodology across diverse datasets and languages for the IC domain. The results in tabular form can be found in Appendix C.

Table 2 finally summarizes the outcomes on LibriSpeech for the ASR task. Similar to the behavior observed for IC, our approach consistently outperforms all baselines, achieving the lowest WER overall (6.32) and among the top-K subgroups (9.33, $K = 5$) and demonstrating superior or comparable results with respect to the two oracles. We observe a clear trend: as we incorporate more data, the performance consistently improves. ASR is inherently more complex than other tasks, such as intent classification. This complexity underscores the significance of our performance improvements. Despite the difficulty of the task, by acquiring only 60% of the entire held-out data, our method achieves performance comparable to using the full dataset. More importantly, our targeted data selection strategy allows for the effective reduction of model biases. For example, we report a top-K WER of 12.24 (with $K = 5$) when all the available data are added (last row of Table 2), whereas our approach achieves a significantly lower top-K WER of 9.33. While our results may not represent the state-of-the-art in ASR, our focus is to demonstrate the effectiveness of the privacy-aware data selection strategy. Specifically, using Whisper base as a model, our approach clearly illustrates how targeted subgroup-based acquisition can enhance performance and mitigate biases effectively.

5 Conclusion

We introduced a data selection strategy to enhance speech model performance while addressing data privacy concerns. Our approach leverages a Challenging Subgroup Identification (CSI) model to detect population subgroups that a model struggles with, without requiring demographic metadata at testing or runtime. We propose acquiring additional data based on the samples labeled as challenging by the CSI model and using them for model retraining. Extensive experiments across two tasks, three datasets, and two languages demonstrate the approach's effectiveness in mitigating biases and outperforming baselines. Its privacy-preserving nature makes it ideal for industry deployment, where collecting demographic data is often restricted. Our results show that the CSI model can be seamlessly integrated into speech recognition pipelines, offering a practical solution for more equitable speech technology in production settings.

Ethical Statement

The paper adheres to the ACL Ethics Policy. This work aims to address fairness and bias in speech recognition systems, which has significant ethical implications. By developing methods that can

mitigate performance disparities without requiring sensitive demographic data, we promote more equitable speech technology while respecting user privacy. However, we acknowledge that any automated system for bias mitigation should be carefully monitored, as it may inadvertently introduce new biases or fail to address all forms of discrimination. Throughout our research and development process, we prioritized transparency, interpretability, and fairness in our methodological choices.

6 Limitations

While our approach shows promising results, a few limitations should be considered. First, the performance of the CSI model depends on the quality and diversity of the initial training data. If certain subgroups are severely underrepresented in the training data, the model may not effectively identify them as challenging. Second, the approach requires a held-out dataset for data selection, which may not always be available in sufficient quantities in real-world scenarios. Finally, computational overhead for training multiple models (speech model, CM, and CSI) may present challenges for resource-constrained deployments. It is worth noting, however, that the CM and CSI models themselves require minimal computational resources, typically converging within minutes. The primary computational costs arise from the two-phase training of the speech model - initial training followed by fine-tuning with the augmented dataset. To address this limitation, future implementations could explore incremental update strategies using parameter-efficient fine-tuning methods such as Low-Rank Adaptation (Hu et al., 2021). These approaches would enable targeted updates to small portions of the model, substantially reducing computational requirements and training time while maintaining performance improvements

Acknowledgements

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.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*.

Arun Babu and et al. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech*.

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

Shihao Chen, Liping Chen, Jie Zhang, KongAik Lee, Zhenhua Ling, and Lirong Dai. 2024. Adversarial speech for voice privacy protection from personalized speech generation. In *ICASSP*.

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*.

Dilek Hakkani-Tür, Giuseppe Riccardi, and Allen Gorin. 2002. Active learning for automatic speech recognition. In *ICASSP*.

Kei Hashimoto, Junichi Yamagishi, and Isao Echizen. 2016. Privacy-preserving sound to degrade automatic speaker verification performance. In *ICASSP*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. arxiv 2021. *arXiv preprint arXiv:2106.09685*.

Thomas Kemp and Alex Waibel. 1998. Unsupervised training of a speech recognizer using tv broadcasts. In *Fifth International Conference on Spoken Language Processing*.

Alkis Koudounas, Flavio Giobergia, Eliana Pastor, and Elena Baralis. 2024a. A Contrastive Learning Approach to Mitigate Bias in Speech Models. In *Proc. Interspeech 2024*.

Alkis Koudounas, Moreno La Quatra, Lorenzo Vaiani, Luca Colomba, Giuseppe Attanasio, Eliana Pastor, Luca Cagliero, and Elena Baralis. 2023a. ITALIC:

An Italian Intent Classification Dataset. In *Proc. Interspeech 2023*, pages 2153–2157.

Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Luca de Alfaro, and Elena Baralis. 2024b. Prioritizing data acquisition for end-to-end speech model improvement. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Luca Cagliero, Luca de Alfaro, Elena Baralis, and Daniele Amberti. 2023b. Exploring subgroup performance in end-to-end speech models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

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. 2024c. Towards comprehensive subgroup performance analysis in speech models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1468–1480.

Alkis Koudounas, Eliana Pastor, Luca de Alfaro, and Elena Baralis. 2025. Mitigating subgroup disparities in speech models: A divergence-aware dual strategy. *IEEE Transactions on Audio, Speech and Language Processing*, 33:883–895.

Alkis Koudounas, Eliana Pastor, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Elisa Reale, Giuseppe Attanasio, Luca Cagliero, Sandro Cumani, Luca de Alfaro, Elena Baralis, and Daniele Amberti. 2024d. Leveraging confidence models for identifying challenging data subgroups in speech models. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*.

Lori Lamel, Jean-Luc Gauvain, and Gilles Adda. 2002. Lightly supervised and unsupervised acoustic model training. *Computer Speech & Language*.

Hui Lin and Jeff A Bilmes. 2009. How to select a good training-data subset for transcription: submodular active selection for sequences. In *Proc. Interspeech*.

Chunxi Liu, Michael Picheny, Leda Sarı, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. 2022. Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions. In *ICASSP*.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding. In *Proc. Interspeech*.

Rishikesh Magar and Amir Barati Farimani. 2023. Learning from mistakes: Sampling strategies to efficiently train machine learning models for material property prediction. *Computational Materials Science*, 224.

M. Padmanabhan, L.R. Bahl, D. Nahamoo, and M.A. Picheny. 1996. Speaker clustering and transformation for speaker adaptation in large-vocabulary speech recognition systems. In *ICASSP*.

Michele Panariello, Francesco Nespoli, Massimiliano Todisco, and Nicholas Evans. 2024. Speaker anonymization using neural audio codec language models. In *ICASSP*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *ICASSP*.

Chanho Park, Rehan Ahmad, and Thomas Hain. 2022. Unsupervised data selection for speech recognition with contrastive loss ratios. In *ICASSP*.

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*, SIGMOD '21. ACM.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*.

Hua Shen, Yuguang Yang, Guoli Sun, Ryan Langman, Eunjung Han, Jasha Droppo, and Andreas Stolcke. 2022. Improving fairness in speaker verification via group-adapted fusion network. In *ICASSP*.

Prakhar Swarup, Roland Maas, Sri Garimella, Sri Harish Mallidi, and Björn Hoffmeister. 2019. Improving ASR Confidence Scores for Alexa Using Acoustic and Hypothesis Embeddings. In *Proc. Interspeech*.

Minh Tran and Mohammad Soleymani. 2023. Privacy-preserving representation learning for speech understanding. In *Proc. Interspeech*.

Irina-Elena Veliche and Pascale Fung. 2023. Improving fairness and robustness in end-to-end speech recognition through unsupervised clustering. In *ICASSP*.

Kai Wei, Yuzong Liu, Katrin Kirchhoff, Chris Bartels, and Jeff Bilmes. 2014a. Submodular subset selection for large-scale speech training data. In *ICASSP*.

Kai Wei, Yuzong Liu, Katrin Kirchhoff, and Jeff Bilmes. 2014b. Unsupervised submodular subset selection for speech data. In *ICASSP*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, and Tim Rault et al. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP: System Demonstrations*.

Yuanyuan Zhang, Yixuan Zhang, Bence Mark Halpern, Tanvina Patel, and Odette Scharenborg. 2022. Mitigating bias against non-native accents. In *Proc. Interspeech*.

A Experimental setup

A.1 Datasets

We evaluate our approach on three publicly available datasets: Fluent Speech Commands (FSC) and ITALIC for the IC task in English and Italian, respectively, and LibriSpeech for ASR. FSC includes 30,043 English utterances, each labeled with three slots (action, object, location) defining the intent. ITALIC consists of 16,521 audio samples from Italian speakers, with the intent defined by action and scenario slots. We select the "Speaker" configuration for ITALIC, aligning with FSC's setup, ensuring distinct speakers in the train, validation, and test sets. For LibriSpeech, we utilize the *clean-360* partition, which comprises 360 hours of clean audio samples. A complete overview of the datasets' characteristics is provided in Table 3.

Metadata. For the above datasets, we consider the following metadata when using DivExplorer to automatically extract subgroups: (i) demographic metadata describing the speaker (e.g., gender, age, language fluency level), (ii) factors related to speaking and recording conditions (e.g., duration of silences, number of words, speaking rate, and noise level), and (iii) intents represented as combinations of action, object, and location for FSC, or action and scenario for ITALIC. We discretize continuous metadata using frequency-based discretization into three distinct ranges, labeled as "low," "medium," and "high". Hence, continuous values are categorized into discrete bins based on their respective frequencies within the dataset. In the experiments, we explore all subgroups with a minimum frequency s of 0.03.

A.2 CM training

We use the following features to train the confidence models:

- *Acoustic embeddings:* We use the embeddings extracted from the audio signal. Specifically, we use the HuggingFace implementation of the wav2vec 2.0 base[2], XLS-R[3], and whisper base[4] models, and we extract the embeddings from the models' last hidden layer.
- *n-best list:* For LibriSpeech, we use the n-best list of the model, i.e., the list of the n most probable hypotheses for each utterance.
- *Output probabilities:* For FSC and ITALIC, we use the output probabilities of the model for each class.
- *Speech metadata:* We use the metadata extracted from the audio signal, including the number of words, number of pauses, speaking rate (word per second), and signal-to-noise ratio.

The CM consists of two hidden layers with GELU activation functions, dropout, and normalization, initialized with the Kaiming normal technique. The CM is trained for up to 10,000 epochs with early stopping, using the NAdam optimizer and a learning rate of 5e-3. For FSC and ITALIC datasets, we use Cross-Entropy (CE) loss. For LibriSpeech, we add a Mean Squared Error (MSE) term, using WER as an additional target. The total loss function is a weighted combination of CE and MSE, defined as: $\mathcal{L}_{tot} = \alpha\mathcal{L}_{CE} + (1-\alpha)\mathcal{L}_{MSE}$, where α is 0.6. The training of the CM takes a few minutes only to converge.

A.3 Models and training procedure

We fine-tune the transformer-based wav2vec 2.0 base (ca. 90M parameters) and multilingual XLSR (ca. 300M parameters) models on the FSC and the ITALIC dataset, respectively, and the whisper base (ca. 74M parameters) model on LibriSpeech. The pre-trained checkpoints of these models are obtained from the Hugging Face hub (Wolf et al., 2020). Experiments were run on a machine equipped with Intel Core ^{TM} i9-10980XE CPU, 2 $\times$ Nvidia RTX A6000 GPU, 128 GB of RAM running Ubuntu 22.04 LTS.

IC task. We trained the models for 2800 steps for FSC and 5100 for ITALIC, with a batch size of 32, using the AdamW optimizer with a learning rate of 1e-4 and 500 warmup steps.

ASR task. We trained the model for 3 epochs, with a batch size of 32, using the AdamW optimizer with a learning rate of 1e-5.

B Subgroups composition

Table 4 presents the top-5 most divergent retrieved subgroups identified by our approach across the three datasets: FSC, ITALIC, and LS. These subgroups represent specific combinations of attributes that exhibit notable performance differences compared to the overall dataset distribution. For the FSC dataset, we observe that subgroups related to

[2] huggingface.co/facebook/wav2vec2-base
[3] huggingface.co/facebook/wav2vec2-xls-r-300m
[4] huggingface.co/openai/whisper-base.en

Table 3: **Datasets characteristics.** Cardinality of the train (`#Train`), held-out (`#Held-out`), validation (`#Val`) and test (`#Test`) sets, the number of distinct speakers (`#Spkr`), and the number of classes (`#C`) for each dataset.

Dataset	#Train	#Held-out	#Val	#Test	#Spkr	#C
FSC (Lugosch et al., 2019)	18506	4626	3118	3793	97	31
ITALIC (Koudounas et al., 2023a)	10498	2625	1957	1441	70	60
LIBRISPEECH (Panayotov et al., 2015)	83211	20803	2703	2620	1001	-

Table 4: **Subgroups composition**. Top-5 most divergent retrieved subgroups for the three considered datasets.

Dataset	Subgroup	Support
FSC	*{action=activate, object=music}*	0.04
	{age=41-65, gender=male, speakRate=high}	0.03
	{gender=male, loc=none, speakRate=high, totSilence=high, trimDur=low}	0.03
	{action=increase, gender=male, nWords=low, speakRate=high}	0.04
	{action=activate, loc=none, speakRate=high, totSilence=high}	0.03
ITALIC	*{gender=male, totSilence=high}*	0.05
	{gender=male, age=22-40, totSilence=high, nWords=low}	0.03
	{speakRate=high, totDur=low, scenario=play}	0.03
	{gender=male, scenario=music, totSilence=high}	0.04
	{nWords=high, nPauses=high, scenario=cooking}	0.03
LS	*{speakRate=high, totDur=low, totSilence=low}*	0.05
	{gender=female, nWords=medium, totDur=high}	0.04
	{nPauses=high, gender=female, totDur=low}	0.03
	{nPauses=low, speakRate=high, totDur=low, totSilence=low}	0.03
	{nPauses=high, nWords=high, speakRate=high}	0.03

voice commands (particularly those involving activation requests and music) demonstrate the highest divergence. Additionally, demographic factors such as male gender combined with high speaking rates appear consistently across multiple subgroups. The ITALIC dataset reveals interesting patterns around specific scenarios, with music-, cooking- and playing-related interactions showing the highest divergence, particularly when combined with male gender and high total silence. In contrast, the LS dataset subgroups are primarily characterized by speech pattern attributes rather than content-based factors. The most divergent subgroup features a high speaking rate combined with low total duration and silence. The female gender appears in two of the top-5 subgroups. These findings highlight the importance of considering fine-grained subgroup performance when evaluating speech recognition systems, as specific combinations of demographic, behavioral, and contextual factors can significantly impact model performance. Most importantly, they highlight the capability of our CSI model to correctly capture demographic information within those subgroups.

C Results on ITALIC

Table 5 presents a comprehensive evaluation of the XLS-R model on the ITALIC dataset, comparing our proposed CSI approach against various baselines and oracle methods. The experiments were conducted across different numbers of challenging subgroups ($K \in [2, 5]$) with corresponding sample selection strategies.

Our CSI method demonstrates superior performance across multiple metrics, consistently achieving the lowest Intent Error Rate (IER) among all non-oracle approaches. For $K = 2$, CSI reduces the IER to 21.94%, which represents a significant improvement over the original model's 26.21%. Notably, this performance is remarkably close to the metadata-based oracle (`M-Oracle`), which achieves 21.12%.

The improvement becomes particularly evident when examining the IER for the top-K most challenging subgroups. CSI reduces the IER top-K from 72.15% in the original model to 59.98% for

Table 5: **ITALIC, XLS-R model.** Mean ±std of three runs. `K` indicates the number of challenging subgroups considered, `N` is the number of samples selected. We compare the results of the `Original` fine-tuning procedure, the baselines, our `CSI`, and the two oracles (`M-Oracle` considering metadata, `S-Oracle` leveraging supervised labels). Best results for each number of subgroups K are highlighted with light-blue. Best results with oracles in **bold**.

K	N	Approach	IER (%) ↓	F1 Macro (%) ↑	IER top-K (%) ↓
-	-	Original	26.21±0.32	68.08±0.37	72.15±0.42 (K = 2)
2	+725	Random	23.95±0.14	72.20±0.19	70.76±0.58
		KNN	23.44±0.06	72.65±0.08	69.13±0.49
		Clustering (Dheram et al., 2022)	22.98±0.14	71.92±0.13	68.05±0.73
		CM	23.70±0.11	71.96±0.08	67.41±0.64
		CSI	21.94±0.10	72.87±0.11	59.98±0.59
		S-Oracle (Magar and Farimani, 2023)	22.86±0.09	72.84±0.12	70.17±0.31
		M-Oracle (Koudounas et al., 2024b)	**21.12±0.12**	**72.94±0.10**	**58.17±0.45**
-	+2625	All data	24.29±0.36	73.22±0.33	65.91±0.34 (K = 2)
3	+975	Random	24.02±0.16	72.01±0.17	66.14±0.64
		KNN	23.59±0.05	71.26±0.09	56.83±0.38
		Clustering (Dheram et al., 2022)	23.17±0.09	71.69±0.08	56.71±0.39
		CM	23.75±0.04	71.88±0.03	57.15±0.55
		CSI	22.50±0.06	72.66±0.04	51.09±0.44
		S-Oracle (Magar and Farimani, 2023)	22.99±0.12	71.77±0.10	57.51±0.42
		M-Oracle (Koudounas et al., 2024b)	**21.74±0.08**	**73.15±0.08**	**50.98±0.38**
4	+1395	Random	23.01±0.11	72.61±0.15	63.94±0.57
		KNN	22.81±0.04	72.48±0.05	55.12±0.37
		Clustering (Dheram et al., 2022)	22.35±0.08	72.78±0.06	55.04±0.29
		CM	22.69±0.05	72.66±0.06	55.61±0.41
		CSI	22.05±0.02	72.86±0.03	49.25±0.43
		S-Oracle (Magar and Farimani, 2023)	22.54±0.07	72.79±0.04	61.02±0.58
		M-Oracle (Koudounas et al., 2024b)	**21.69±0.03**	**73.24±0.04**	**47.16±0.19**
5	+1509	Random	23.59±0.15	72.26±0.17	58.49±0.71
		KNN	23.09±0.04	72.04±0.04	48.15±0.48
		Clustering (Dheram et al., 2022)	22.19±0.02	72.85±0.03	50.71±0.22
		CM	23.10±0.05	71.99±0.04	49.74±0.43
		CSI	22.14±0.01	72.30±0.03	42.19±0.39
		S-Oracle (Magar and Farimani, 2023)	22.56±0.03	72.85±0.05	60.56±0.19
		M-Oracle (Koudounas et al., 2024b)	**21.95±0.04**	**72.99±0.05**	**41.88±0.21**
-	+2625	All data	24.29±0.36	73.22±0.33	58.44±0.37 (K = 5)

$K = 2$ and achieves an even more important reduction to 42.19% for $K = 5$. This represents an improvement of approximately 17% and 42%, respectively, demonstrating CSI's effectiveness in addressing performance disparities.

Furthermore, CSI consistently outperforms established baselines, including Random sampling, KNN, Clustering, and CM approach across all K values. The performance gap is particularly pronounced for the IER top-K metric, indicating CSI's superior ability to target and improve model performance on the most challenging subgroups.

Interestingly, CSI's performance closely approximates the `M-Oracle`, which leverages sensitive demographic metadata. This suggests our approach can effectively identify and address performance disparities without requiring direct access to potentially sensitive attributes like age and gender. For $K = 5$, CSI achieves an IER top-K of 42.19%, nearly matching `M-Oracle`'s 41.88%.

When compared to the supervised oracle (`S-Oracle`), which utilizes ground truth labels, CSI demonstrates superior performance on the IER top-K metric across all K values. This highlights CSI's advantage in specifically addressing subgroup disparities rather than simply focusing on overall error reduction.

These results confirm that our CSI approach effectively identifies challenging subgroups, strategically selects additional training samples, and significantly improves model fairness and overall performance without requiring access to sensitive attributes or supervised labels.