Mitigating Subgroup Disparities in Speech Models: A Divergence-Aware Dual Strategy

Alkis Koudounas *Graduate Student Member, IEEE,* Eliana Pastor, Luca de Alfaro, Elena Baralis *Member, IEEE*

***Abstract*—Speech models may exhibit disparities in performance across different population subgroups. Prior mitigation efforts often rely on the manual (user-driven) selection of predefined data subgroups of interest. However, these approaches may fail to correctly identify all relevant subgroups associated with performance issues.**

We propose a dual strategy to mitigate subgroup performance disparities by automatically identifying the subgroups showing a worse performance, i.e., a *divergence*, compared to the overall model performance. We consider diverse metadata, such as speaker gender and utterance duration, to delineate the subgroups and enable their interpretation. Once the subgroups are identified, we tackle the performance disparities from two alternative perspectives - a post-processing, refining pre-trained models after fine-tuning, and an in-processing one, implementing mitigation measures during model development. In the post-processing strategy, we propose a divergence-aware data acquisition to prioritize acquiring samples from underperforming subgroups. The in-processing scenario introduces two approaches: a divergence-based regularization and a data augmentation technique to boost subgroup performance during model fine-tuning. Experiments on a dataset for Automatic Speech Recognition, one for Emotion Recognition, and two datasets for Intent Classification in English and Italian highlight the improvement achieved by the divergence-aware strategies, which significantly reduce performance disparities and outperform traditional clustering-, KNN- and random-based methods.

***Index Terms*—bias mitigation, spoken language understanding, speech processing, data acquisition, divergence**

I. Introduction

THE advances in speech and language technologies have transformed how we interact with machines over the past few years. These innovations have become ubiquitous in our daily lives, from voice-activated virtual assistants to language translation tools. However, as these models become more pervasive, there is also a growing concern about potential disparities in their behavior. Biases present in training data, linguistic variations, and disparate data representation can inadvertently lead to unequal outcomes, affecting certain subgroup populations more than others. Recent studies revealed model bias and disparate treatment in data subgroups ([1]–[7]), emphasizing the need for addressing these issues. Identifying and mitigating these disparities is crucial to ensure that speech and language technologies are fair and robust across subpopulations.

Current mitigation solutions often rely on a priori knowledge or user-driven selection of the subgroups of interest. These strategies primarily focus on the diversity and robustness of the data, addressing challenges related to linguistic variations, recording conditions, environment, and demographics [8]. However, these approaches may overlook unexpected subgroups associated with performance issues. Moreover, disparities may emerge at the intersection of multiple challenging characteristics. Recent advancements in mitigation solutions identify data subgroups automatically [1]. However, the identified subgroups lack interpretability, not allowing the interpretation of the source of disparities. Consequently, these approaches cannot guide targeted data acquisition to mitigate disparities.

In this paper, we propose to automatically identify interpretable data subgroups that exhibit a performance disparity with respect to the overall model behavior, denoted as *divergence*, and subsequently mitigate the performance disparities within the identified subgroups. For the automatic identification, we leverage the techniques of [9], [10] that extract data subgroups with divergent behavior in performance. The subgroups are defined as interpretable combinations of metadata such as speaker demographics, recording conditions, and task characteristics.

To mitigate disparities, we propose two alternative strategies—post-processing, refining the performance of a pre-trained model after fine-tuning for the downstream task, and in-processing, implementing mitigation measures during the model development.

As a post-processing strategy, we propose a divergence-aware data acquisition. Given a trained model, we leverage the identified subgroups with worse performance than the average to guide the data acquisition process. Being the subgroup interpretable, we can, for example, reveal that our model struggles with utterances of women speaking fast, and we acquire samples with such metadata.

In the in-processing scenario, we tackle the mitigation process during the model development. We propose two methods: divergence-aware regularization and divergence-aware data acquisition. As for regularization, we introduce a novel regularization term directly associated with the divergence of each subgroup. This term emphasizes subgroups showing a more pronounced performance disparity. During model training, samples belonging to subgroups with higher divergence, i.e., greater differences from overall model performance, receive greater attention. For data augmentation, we augment audio data from subgroups the model struggles with by applying diverse transformations. This subgroup-based data augmentation increases the representation of such challenging samples, thus improving model robustness and performance at the subgroup level.

The dual-strategy, in-processing or post-processing, offers versatility to accommodate diverse application needs, usability

AK, EP, and EB are with the Politecnico di Torino, Turin, Italy.
LdA is with the University of California, Santa Cruz, CA, USA.

constraints, and purposes. Practitioners can opt for one strategy or the other based on their specific requirements, whether it be improving a trained model, the feasibility of collecting additional samples, or the goal of directly training a subgroup-regularized model.

We evaluated our dual approach on the LIBRISPEECH [11] dataset for Automatic Speech Recognition (ASR), IEMOCAP [12] for Emotion Recognition (ER), and two Spoken Language Understanding (SLU) datasets for intent classification, FSC [13] for the English language and ITALIC [14] for Italian. We employ the transformer-based wav2vec 2.0 [15] model for the IC and ER English datasets, the multilingual XLSR [16] model for ITALIC, and Whisper [17] for LIBRISPEECH. The experimental findings underscore the efficacy of our approaches in mitigating performance disparities. Specifically, our post-processing demonstrates that targeted sample acquisition improves subgroups and overall model performance compared to existing clustering-based, KNN, and error-based baselines and indiscriminate data acquisition. In the case of our in-processing techniques, we show their ability to reduce disparities during the training process, with the regularization slightly outperforming the subgroup-based data augmentation, enabling the direct development of models with enhanced fairness and equity in the outcomes.

Our main contributions are the following.

- We address the challenge of mitigating subgroup disparities by introducing a divergence-aware dual strategy that leverages the automatic identification of divergent data subgroups. Our approach encompasses both in-processing and post-processing methods, allowing practitioners to adapt to diverse requirements and constraints in real-world scenarios.
- We introduce a post-processing technique that mitigates disparities in data subgroups via a divergence-aware data acquisition.
- We propose two in-processing mitigation approaches to boost the performance of divergent subgroups through regularization or data augmentation during the training process.

We introduced a preliminary version of this work in [18], which focused only on the divergence-aware data acquisition process and intent classification tasks. This paper proposes a more thorough approach to mitigating subgroup disparities by introducing also two in-processing techniques. In-process mitigation allows for reducing disparities even when data acquisition campaigns may not be optimal or feasible. This expanded framework provides a more flexible solution to address subgroup disparities across various practical scenarios, enhancing its applicability.

The source code and its documentation to adopt our approach and reproduce the results are available at `https://github.com/koudounasalkis/DADS`.

We organize the paper as follows. Section II reviews the related works. Section III describes our dual strategy. Section IV presents the experimental setting. Section V reports the main experimental results. Finally, Section VI draws the conclusions.

II. RELATED WORK

The increasing use of speech systems has raised concerns about potential biases, leading to various recent studies exploring different aspects of bias and disparity [2], [4], [19]–[24]. These works generally address the challenges related to linguistic variations, recording conditions, environment, and demographics [8]. Several works have examined racial bias [2], [19], performance disparities across gender, dialects, and race [20] or age [4], and the impact of gender representation in speech corpora [21], [22]. Studies have also questioned conventional evaluation metrics [23] and introduced corpora [24] to identify demographic bias in speech applications.

Some techniques propose to address disparities during the training process, such as via domain adversarial training [25], or counterfactual modifications of dependent variables, such as the speaker's voice [26]. Privacy-preserving techniques that extract utterance level embeddings using a speaker ID model and group embedding adaptation have also been explored for fairness improvement in Automatic Speech Recognition (ASR) and Speaker Verification systems [27], [28]. The considered groups are user-defined and known a priori. In contrast, our approach can automatically identify the subgroup where the model behaves differently and mitigate such disparities. As a result, we boost overall and subgroup-level performance.

Recent work on acquisition-based techniques addressed the question of how many data samples we should acquire from each group to improve model performance using learning curves [29], [30], given a set of groups of interest. The work in [25] also explores data augmentation for known critical subgroups, such as non-native speakers, to augment training data. Closer to our work, the approach in [1] automatically groups data by clustering speaker embeddings and identifies the clusters that exhibit inferior performance for a given model. Data acquisition considers data samples close to the problematic clusters. However, these clusters, unlike our subgroups, are not interpretable. Subgroup interpretability allows for guiding the data acquisition process, collecting data with specific properties. Non-interpretable subgroups instead only allow selecting from already-available data, e.g., by feeding it into an encoder model that extracts embeddings.

The work in [31] identifies subgroups defined by attribute combinations but concentrates on under-represented subgroups. We instead focus on identifying all subgroups with adequate representation in data on which the model *underperforms*, be it for a trained model or during the training process, so that we can acquire more data or boost the model to address this issue specifically. Combining these approaches could offer a two-fold solution for bias mitigation.

III. A DUAL STRATEGY FOR MITIGATION

Our approach to mitigating bias and improving fairness in a model, either already trained or during training [1], involves two main steps: (i) automatic identification of subgroups

[1]In our work, we fine-tune the pre-trained self-supervised models. In the rest of the paper, we will use both training and fine-tuning without distinction to refer to the fine-tuning process.

(Section III-A) and (ii) divergence-aware mitigation. In the first step, we extract interpretable subgroups and compute how the model performs on these subgroups compared to its overall performance. The second step involves either post-processing through targeted data acquisition (Section III-B) or in-processing mitigation III-C), depending on whether we consider an already trained model or actively training one.

A. Automatic Subgroup Identification

Consider a dataset of utterances D. We annotate each utterance with a set of interpretable metadata. These can be speaker-related features, such as gender or age, or speaking and recording features, such as utterance duration, presence of noise, and speaking rate. The metadata can be either already available in the dataset, such as the self-declared gender and age of the speaker, or automatically derived [9] from utterances, such as the speaking rates as words per second. A data subgroup S is a subset of the dataset D sharing the same set of metadata. We represent a subgroup as a conjunction of attribute-value pairs. For example, the subgroup {gender=female, duration>5s} represents utterances of female speakers with a duration greater than 5s.

Consider a model M and a subgroup S. $f(S, \mathcal{M})$ denotes a performance measure (e.g., accuracy) of M on subgroup S. The *divergence* [10] of subgroup S for model M and measure f, denoted $\Delta_f(S, \mathcal{M})$, quantifies the difference between the model performance on subgroup S and its performance on the entire dataset D:

$$\Delta_f(S, \mathcal{M}) = f(S, \mathcal{M}) - f(D, \mathcal{M}) \ . \tag{1}$$

The higher the divergence, the more its performance diverges from the overall one. For example, a high negative divergence in accuracy for a subgroup indicates the model struggles with the utterances of that subgroup.

We adopt the identification procedure described in [9] to derive metadata, extract subgroups, and compute their divergence. Specifically, we use the DIVEXPLORER [10], [32] approach, which identifies all subgroups with an adequate representation in the dataset based on a frequency threshold, denoted minimum support $minsup$. The support threshold $minsup$ (such as 0.1% of the dataset) controls the exploration and ensures that the subgroups contain enough utterances to make the performance computation statistically significant. This is critical as performance measures on subgroups with small support can be subject to statistical fluctuations. The resulting subgroups, denoted as *frequent*, can overlap. For example, the subgroup {gender=female, duration>5s} overlaps with {gender=female}.

In summary, given a dataset with annotated metadata, a model M, and a performance measure of interest f, we identify the set of frequent subgroups $\mathcal{S}$. For each $S \in \mathcal{S}$, we have its divergence $\Delta_f(S, M)$, and the statistical significance t of the divergence computed with the Welch t-test. This information about divergent subgroups the model struggles with enables us to actively address these issues, either post-processing through targeted data acquisition (Section III-B) or in-processing via regularization or data augmentation (Section III-C).

B. Post-processing mitigation

Post-processing mitigation involves mitigating subgroup disparities of an already trained model. We propose to mitigate disparities by acquiring data from subgroups the model struggles with and retraining it accordingly. In the following, we describe this methodology.

***Divergence-aware Data Acquisition**:* Let $\mathcal{S}$ be the set of frequent subgroups, and each subgroup $S \in \mathcal{S}$ is characterized by divergence $\Delta_f(S, \mathcal{M})$ for the performance measure f. We define $\mathcal{S}^- \subseteq \mathcal{S}$ as the set of *challenging* subgroups for which model $\mathcal{M}$ has lower performance than the average. For performance measures for which the higher, the better (e.g., accuracy, F1 measure), $\mathcal{S}^-$ consists of the subgroups that have negative divergence, i.e., $\mathcal{S}^- = \{S \in \mathcal{S} | \Delta_f(S) < 0\}$. We can easily modify this definition to apply to the opposite case (e.g., word error rate, the lower, the better).

We perform a pruning step to reduce redundancy among the challenging subgroups, following the pruning approach outlined in [10]. During this pruning process, when presented with two subgroups, S_a and S_b, where S_b includes S_a along with an additional metadata condition, we retain only the more general S_a if the absolute difference in the divergence between the two subgroups falls below a predefined threshold. The rationale behind this approach is that S_a already represents the divergence of S_b, as the additional metadata of S_b only marginally affects the divergence. For instance, consider the subgroup {*young_woman*} with a divergence of -0.39 and {*young_woman, utterance_duration>10s*} with a divergence of -0.41. In this scenario, we preserve solely the former subgroup, as it accounts for most of the divergence observed in the latter. We denote the summarized set with $\dot{\mathcal{S}}^-$. Pruning the challenging subgroups results in a more concise representation and facilitates data acquisition, as we can focus on the most relevant subgroups.

Our subgroups are *interpretable*. Hence, we can specifically target the acquisition of data samples with characteristics of the identified challenging subgroups.

We target for performance improvement the top-K summarized challenging subgroups $\dot{\mathcal{S}}^-$ with the highest absolute divergence by acquiring data belonging to these subgroups and denote them with $\dot{\mathcal{S}}_k^-$. This selection of only the most divergent challenging subgroups allows us to focus and control the targeted acquisition process.

Once we identify $\dot{\mathcal{S}}_k^-$, the mitigation process via data acquisition is straightforward. Specifically, we retrain the model by adding new data belonging to one or more of the K subgroups (as subgroups can partially overlap, the same data instance can belong to more than one top-K subgroup).

More formally, let $\mathcal{T}$ be the training set and $\mathcal{U}$ a set of utterances unseen at training time. Utterance $x_i \in \mathcal{U}$ satisfies a subgroup S, denoted as $x_i \vdash S$, if its metadata values match S. The data acquisition consists of acquiring a set of new utterances $\mathcal{U}(\dot{\mathcal{S}}_k^-)$ satisfying at least a challenging group $\dot{\mathcal{S}}_k^-$, with $\mathcal{U}(\dot{\mathcal{S}}_k^-) = \{x_i \in \mathcal{U} \mid \exists S \in \dot{\mathcal{S}}_k^- : x_i \vdash S\}$. Finally, the mitigation step consists of retraining the entire model $\mathcal{M}$ of the enriched dataset $\mathcal{T} \cup \mathcal{U}(\dot{\mathcal{S}}_k^-)$. The parameter K allows us to control the data acquisition process. Our experiments

will illustrate how the choice of K affects overall model performance as well as subgroup-specific performance.

C. In-processing mitigation

In-processing mitigation involves addressing disparities in data subgroups during model training. We propose to use the information of the subgroups the model struggles with and their divergence for improving the model by operating either (i) on the model loss or (ii) on the data themselves.

For the former approach, we introduce a regularization term into the model loss. This term encourages the model to focus more on data samples from subgroups where performance diverges from the model's overall behavior. The regularization strength is proportional to the extent of this divergence. The latter approach involves data augmentation for samples within struggling subgroups. By enriching the dataset with augmented versions of these samples, we aim to improve the model's ability to handle such challenging subgroups.

Divergence-Aware Regularization: We propose a divergence-based regularization term to mitigate subgroup disparities at training time. At each epoch, we derive subgroup divergence scores to guide the training process accordingly. Intuitively, the higher the divergence of a subgroup, the more the model deviates in modeling it compared to the overall data. Consequently, the model should focus on the data samples belonging to this subgroup to mitigate its divergent behavior. Essentially, the regularization terms encourage the model to adjust its focus based on the degree of divergence, prioritizing data samples that belong to challenging subgroups.

Let $\mathcal{T}$ and $\mathcal{V}$ be the training and validation sets. Given a model $\mathcal{M}_e$ at epoch e trained on $\mathcal{T}$, we extract the set $\mathcal{S}$ of frequent subgroups coupled with their divergence scores from the validation set $\mathcal{V}$. Let x_i be an utterance in $\mathcal{T}$, and y_i and $\hat{y}_i$ its true and predicted labels. We denote by $\mathcal{S}(x_i)$ the set of subgroups satified by x_i, with $\mathcal{S}(x_i) = \{S \in \mathcal{S} \mid x_i \vdash S\}$.

For each utterance x_i, we define a boosting weight equal to the highest absolute divergence among the subgroups satisfied by x_i:

$$w(x_i) = \max_{S \in \mathcal{S}(x_i)} |\Delta_f(S, \mathcal{M})| \tag{2}$$

We introduce the *divergence-based* loss L_Δ:

$$\mathcal{L}_\Delta = \sum_{x_i \in \mathcal{T}} w(x_i) \mathcal{L}_{CE}(y_i, \hat{y}_i) \tag{3}$$

where L_{CE} denotes the standard cross-entropy loss. Utterances associated with higher divergence will have a greater impact on the divergence-based loss $\mathcal{L}_\Delta$. The final loss is defined as

$$\mathcal{L} = \alpha \mathcal{L}_{CE} + (1 - \alpha) \mathcal{L}_\Delta \tag{4}$$

where $\mathcal{L}_\Delta$ is the regularization term and α is weighting factor. The $\mathcal{L}_\Delta$ term allows the model to give more attention to utterances that exhibit larger divergences.

Algorithm 1 summarizes each step of our training strategy. We first initialize the boosting weights (Line 1) and extract metadata from the utterances in the training and validation sets (Line 2). Then, the training procedure iterates the following steps for all epochs. (i) Train the model with the training loss defined in Equation 4, (ii) Extract the subgroups with a frequency greater than s and their divergence scores using DIVEXPLORER (Line 5), (iii) For each utterance x_i in training set $\mathcal{T}$, derive the set of subgroups satisfying it, i.e., $\mathcal{S}(x_i)$ (Line 6), (iv) update the boosting weights w for each training instance via Equation 2 (Line 7). Finally, the algorithm returns the final boosted model (Line 9).

Algorithm 1 Divergence-Aware Regularization

Require: Training Set $\mathcal{T}$, Validation Set $\mathcal{V}$, min frequency s
Ensure: M: Model

1: Initialize weights $w(x_i) = 1.0 \ \forall x_i \in \mathcal{T}$
2: $\mathcal{T}_m, \mathcal{V}_m \leftarrow$ Derive metadata $\mathcal{T}, \mathcal{V}$
3: **for each** epoch $e \in E$ **do**
4: $\quad \mathcal{M}_e \leftarrow$ Model trained on $\mathcal{T}$ at epoch e via Eq. 4
5: $\quad \mathcal{S}$, $\Delta_f(S, \mathcal{M}_e) \forall S \in \mathcal{S} \leftarrow$ DIVEXPLORER($\mathcal{M}_e, \mathcal{V}_m$)
6: $\quad \mathcal{S}(x_i) \leftarrow$ Satisfy($x_i, \mathcal{S}$) $\quad \forall x_i \in \mathcal{T}_m$
7: $\quad w(x_i) \leftarrow$ ComputeWeights($x_i, \mathcal{S}(x_i)$) $\quad \forall x_i \in \mathcal{T}_m$ via Eq. 2
8: **end for**
9: **return** $\mathcal{M}_e$

Divergence-Aware Data Augmentation: While the regularization strategy adjusts the training process at the subgroup level by modifying the loss function, the data augmentation strategy operates on the data itself. In summary, at each epoch, we derive the subgroups the model mostly struggles with. We perform data augmentation on the training samples satisfying at least one of those challenging subgroups and keep training the model on such an augmented dataset. Intuitively, the model can better learn such critical cases and improve performance by augmenting challenging samples.

More formally, being $\mathcal{V}$ the validation set and $\mathcal{M}_e$ the model at epoch e, we compute the top-K summarized challenging subgroups $\dot{\mathcal{S}}^-$ with the highest absolute divergence on $\mathcal{V}$ and for model $\mathcal{M}_e$. $\dot{\mathcal{S}}^-$, derived likewise to the post-processing technique, are the summarized subgroups the model most struggles with. Being $\mathcal{T}_b$ the batch of training data, we consider the set of utterances $\mathcal{T}_b(\dot{\mathcal{S}}_k^-)$ satisfying at least a challenging group $\dot{\mathcal{S}}_k^-$, with $\mathcal{T}_b(\dot{\mathcal{S}}_k^-) = \{x_i \in \mathcal{U} \mid \exists S \in \dot{\mathcal{S}}_k^- : x_i \vdash S\}$. Again, this resembles the post-processing strategy, but in this case, the set is from the training rather than an unseen set. We then perform data augmentation on samples $\mathcal{T}_b(\dot{\mathcal{S}}_k^-)$. Data augmentation techniques include time stretching, background noise injection, reverberation, pitch shifting, or a random combination of these perturbations. Hence, the mitigation step consists of training the entire model $\mathcal{M}$ of such an augmented set. By augmenting these challenging samples, we provide the model with additional training instances that can help it better model the challenging subgroups.

IV. EXPERIMENTAL SETTING

A. Datasets

We evaluated our approach on three tasks: Intent Classification (IC), Emotion Recognition (ER), and Automatic Speech Recognition (ASR), focusing on datasets in both English and

Italian languages. Specifically, we considered the following four datasets.

FLUENT SPEECH COMMANDS (FSC) [33] is a dataset in English for the IC task, including 30,043 utterances from 97 speakers. Each audio sample has three slots: action, object, and location, determining 31 distinct intents. We used the intent accuracy as target performance f.

ITALIC [14] is a dataset for IC in Italian containing 16,521 samples from 70 speakers. The action and scenario slots denote the intents for a total of 60 distinct intents. We considered the intent accuracy.

LIBRISPEECH [11] is a collection of audio recordings from audiobooks. We use the *"clean-360"* version, which includes 360 hours of clean audio samples. We evaluated the Word Error Rate (WER) performance metric for the ASR task.

IEMOCAP [12] - Interactive Emotional Dyadic Motion Capture is a dataset for the ER task. The dataset includes discrete emotion labels (i.e., happiness, anger, sadness, frustration, and neutral state) and continuous arousal annotations (i.e., activation, valence, and dominance). Following standard procedure [34], we considered four classes (neutral, happy, sad, angry) to have balanced emotion categories, resulting in a dataset of 4,990 samples. We used the emotion label accuracy as a target performance measure.

For FSC, ITALIC, and LIBRISPEECH datasets, we considered the official splits of train, validation, and test sets, with each speaker exclusively assigned to one set. The IEMOCAP dataset is divided into five sessions (i.e., splits) typically evaluated using a 5-fold cross-validation approach. In this study, we used three sessions as the training set, one as validation, and one as test set to match the configuration of the other datasets. As a result, we have the training, validation, and test sets for all four datasets. We evaluated the proposed mitigation techniques in two configurations. In the first configuration, we use the full train set for training the speech model. We then identify the frequent subgroups, their divergence, and the subset of challenging subgroups on the validation set. Finally, we evaluated the model performance on the test set. We adopted this configuration exclusively for the in-processing techniques as the post-processing data acquisition requires unseen labeled data. In the second configuration, we partitioned the training set, allocating 80% for training and holding out 20%. We used the held-out set to acquire data samples for the post-processing technique. We used the validation and test sets in the same way as the first configuration, maintaining consistency across all techniques. We use this configuration for all techniques, both post- and in-processing.

B. Metadata

We implemented the metadata enrichment as proposed in [35], considering demographic information, speaking and recording conditions, and dataset-specific metadata. Regarding speaker demographics, we considered all available (self-declared) demographic data, such as age, gender, and country of origin. For speech-oriented metadata, we derived the number and the duration of silence (both total and trimmed), the word count, and the speaking rate (words per second). As dataset-dependent metadata, we analyzed metadata specific to each dataset and/or task. We used the intent slots for the FSC and ITALIC datasets and the emotion and arousal annotations for IEMOCAP. Continuous metadata was discretized into "low", "medium", and "high" frequency bins.

C. Models

We fine-tuned the pre-trained wav2vec 2.0 [15] base model for the FSC and IEMOCAP datasets, the multilingual XLSR [16] model for ITALIC, and Whisper [17] base monolingual for LIBRISPEECH. We used the pre-trained checkpoints available on the Hugging Face hub [36].

D. Hypeparameter setting

In our subgroup extraction with DivExplorer, we explored all subgroups with a minimum frequency of 0.03, following [18]. The α weighting parameter for the regularization loss is set to 0.7 For both post-processing data acquisition and in-processing data augmentation, the hyperparameter K defines the top-K most challenging subgroups. We varied the value of K from 2 to 5, and we analyzed its impact on the results. For the core of the experiments, we use K=2 for both the data acquisition and targeted data augmentation as it has been shown to lead to the best results overall [18]. Specifically, K=2 corresponds to 226 additional samples for FSC, 154 for ITALIC, 112 for IEMOCAP, 6715 for LIBRISPEECH. We then studied the impact of varying K for a sensitivity analysis. Our fine-tuning process included a hyperparameter search and followed established procedures outlined in the relevant literature. Each IC and ER model undergoes fine-tuning by adding a final classification layer to the encoder architecture. Specifically, for IC and ER, we utilized a learning rate of 1e-4, a batch size of 32, a warmup ratio of 0.1, and a weight decay of 0.01. In the case of ASR, we fine-tuned the entire Whisper base model, employing a learning rate of 1e-5, a batch size of 8, 500 warmup steps, and a weight decay of 0.01. For all models, we opted for the AdamW optimizer. The IC and ER models were trained for a maximum of 30 epochs with an early stopping criterion, while the ASR model underwent a maximum of 5 epochs of training. Experiments were run on a machine equipped with Intel® Core™ i9-10980XE CPU, $1 \times$ Nvidia® RTX A6000 GPU, 64 GB of RAM running Ubuntu 22.04 LTS.

E. Metrics

We evaluated the overall model performance using accuracy and macro F1 scored for FSC, ITALIC, and IEMOCAP and *WER* (Word Error Rate) and *CER* (Character Error Rate) for LIBRISPEECH. We also assessed the performance at the subgroup level. We focused on the most challenging subgroup, i.e., the subgroup that shows the most substantial decrease in performance compared to the overall average, denoted with Δ^-_{max}. Δ^-_{max} evaluates how well the model can reduce differences in performance and thus mitigate bias. We also computed the average divergence on the top 10, 20, and 50 subgroups with the highest decrease in performance (Δ^-_{avg-n}

TABLE I

MEAN AND STANDARD DEVIATION RESULTS OF THREE RUNS ON THE CONSIDERED IC DATASETS. ORIGINAL FINE-TUNING AND MITIGATION STRATEGIES, INCLUDING ACQUISITION, REGULARIZATION, AND TARGETED DATA AUGMENTATION (TARGET DATA++), CONSIDERING THE ORIGINAL TRAINING SET DIVIDED INTO TRAINING AND HELD-OUT SETS, K=2. BEST RESULTS FOR EACH DATASET ARE IN **BOLD**, SECOND-BEST UNDERLINED; BEST RESULTS FOR EACH DATASET AND STRATEGY IN LIGHT YELLOW.

DS	*Method*	*Strategy*	*Accuracy*	*F1 Macro*	Δ^-_{max}	Δ^-_{avg-10}	Δ^-_{avg-20}	Δ^-_{avg-50}	$\lvert\Delta_{avg-all}\rvert$
FSC	original	-	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
	w/ random	acquisition	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
	w/ KNN	acquisition	92.07±0.17	89.92±0.11	-49.90±0.33	-49.85±0.29	-49.76±0.27	-46.98±0.28	0.96±0.03
	w/ clustering	acquisition	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
	w/ supervision	acquisition	95.71±0.74	94.06±0.83	-48.13±0.39	-48.02±0.36	-47.58±0.35	-45.97±0.48	0.92±0.04
	ours	acquisition	**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
	w/ random	target data++	92.85±0.75	92.29±0.68	-45.67±2.78	-45.59±2.75	-43.41±2.68	-41.28±2.51	0.84±0.27
	w/ KNN	target data++	93.94±0.28	93.15±0.31	-43.61±1.32	-43.34±1.24	-42.12±1.19	-38.84±1.08	0.75±0.03
	w/ clustering	target data++	94.49±0.41	94.31±0.44	-40.09±2.12	-39.95±2.03	-39.77±1.84	-34.65±1.07	0.38±0.10
	ours	target data++	95.75±0.37	95.48±0.35	-36.12±0.39	-35.98±0.37	-34.77±0.36	-32.65±0.33	0.35±0.04
	w/ random	regularization	93.41±0.52	93.22±0.67	-44.51±6.59	-44.25±6.55	-44.04±6.21	-38.54±5.85	0.85±0.14
	w/ KNN	regularization	95.11±0.21	95.04±0.20	-41.32±3.52	-41.19±3.28	-40.51±3.15	-36.95±2.75	0.62±0.05
	w/ clustering	regularization	95.75±0.39	95.49±0.41	-39.51±5.68	-39.18±5.21	-37.29±4.74	-34.74±4.18	0.43±0.02
	ours	regularization	96.47±0.11	**96.33±0.12**	**-34.49±0.45**	**-34.49±0.45**	**-34.11±0.41**	**-31.34±0.32**	**0.29±0.01**
	original	all data	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
ITALIC	original	-	73.79±0.32	68.08±0.37	-47.63±1.93	-47.52±1.94	-47.15±1.92	-43.31±1.78	0.60±0.01
	w/ random	acquisition	75.32±0.63	70.72±0.58	-47.00±0.81	-46.86±0.80	-46.22±0.77	-42.68±0.70	0.48±0.02
	w/ KNN	acquisition	75.56±0.57	70.21±0.54	-46.11±0.93	-46.02±0.92	-45.49±0.84	-42.17±0.74	0.39±0.02
	w/ clustering	acquisition	74.05±0.33	69.09±0.75	-45.02±2.02	-44.91±2.01	-44.14±1.81	-39.79±1.33	0.37±0.08
	w/ supervision	acquisition	77.14±0.52	72.65±0.63	-46.97±1.15	-46.84±1.07	-45.91±1.02	-42.36±0.93	0.45±0.04
	ours	acquisition	**77.40±0.24**	72.51±0.14	-31.75±0.55	-31.71±0.55	-31.11±0.41	-28.19±0.18	0.34±0.03
	w/ random	target data++	75.14±0.49	73.01±0.79	-46.89±2.05	-46.51±1.98	-44.98±1.57	-42.04±1.36	0.35±0.12
	w/ KNN	target data++	75.97±0.34	73.67±0.39	-41.19±1.17	-40.53±1.06	-38.57±0.95	-35.77±0.89	0.31±0.03
	w/ clustering	target data++	76.59±0.84	73.98±0.78	-38.95±2.69	-38.37±2.43	-37.01±2.20	-34.15±2.02	0.28±0.04
	ours	target data++	77.12±0.54	74.05±0.42	-31.93±1.91	-31.58±1.85	-30.05±1.59	-28.19±1.35	0.23±0.05
	w/ random	regularization	76.04±0.71	72.11±0.55	-46.58±2.29	-46.22±2.21	-45.87±2.08	-43.16±1.97	0.33±0.11
	w/ KNN	regularization	76.54±0.44	73.08±0.39	-41.23±1.24	-41.04±1.17	-38.63±1.02	-35.78±0.87	0.29±0.04
	w/ clustering	regularization	76.67±0.79	74.01±0.76	-38.43±2.51	-38.05±2.18	-36.59±1.96	-33.93±1.79	0.25±0.03
	ours	regularization	77.02±0.61	**74.19±0.48**	**-31.54±2.02**	**-31.14±1.93**	**-29.88±1.74**	**-28.10±1.67**	0.21±0.05
	original	all data	75.71±0.36	73.22±0.33	-47.54±0.79	-47.36±0.76	-46.68±0.47	-41.93±0.00	**0.15±0.03**

), along with the average absolute divergence across all identified subgroups ($\lvert\ \Delta^-_{avg-all}\ \rvert$). Note that, for LIBRISPEECH, a subgroup's poorer performance compared to the overall system is reflected by a larger divergence in its WER value. Therefore, unlike the divergence in accuracy for the other datasets, a positive WER divergence signifies reduced performance. These metrics enable us to quantify the effectiveness of the mitigation approach in addressing performance discrepancies across subgroups.

F. Baselines

We benchmark our in- and post-processing mitigation approaches against four alternative approaches to derive challenging samples to mitigate.

Random baseline. As a straightforward benchmark, we randomly select the challenging samples. This approach serves as a baseline for comparison and to demonstrate the need for subgroup-based and divergent-aware selection.

Cluster-based baseline [1]. We identify the challenging subgroups via unsupervised clustering, following the approach proposed in [1]. We first extract acoustic embeddings from audio samples in the validation set. We apply K-means clustering to group them into similar clusters. Following [1], we used 50 clusters for LIBRISPEECH as they are proven to adequately capture speech characteristics pertinent to ASR. We instead considered 10 clusters for ITALIC and 20 for FSC, as these configurations have been found to achieve the best performance on the target datasets [18]. For IEMOCAP, we also examined 20 clusters as this configuration led to the best results overall. We then select clusters with the poorest performance, representing the challenging subgroups the model struggles with. We finally take challenging samples based on their proximity to these identified subgroups.

KNN baseline. We employ a K-Nearest Neighbors (KNN) technique. We assess whether an utterance is challenging for the model or not by conducting a majority vote among its neighbors in the training set, considering instances where the model incorrectly classifies them. K is chosen by optimizing performance, i.e., identifying challenging subgroups, on the validation set. Specifically, we use K equal to 14 for FSC, 11 for ITALIC, 12 for IEMOCAP, and 18 for LIBRISPEECH.

Error-driven baseline [37]. We adopt an error-driven ap-

TABLE II
MEAN AND STANDARD DEVIATION RESULTS OF THREE RUNS ON IEMOCAP AND LIBRISPEECH DATASETS. ORIGINAL FINE-TUNING AND MITIGATION STRATEGIES, INCLUDING ACQUISITION, REGULARIZATION, AND TARGETED DATA AUGMENTATION (TARGET DATA++), CONSIDERING THE ORIGINAL TRAINING SET DIVIDED INTO TRAINING AND HELD-OUT SETS; K=2. BEST RESULTS FOR EACH DATASET ARE IN **BOLD**, SECOND-BEST UNDERLINED; BEST RESULTS FOR EACH DATASET AND STRATEGY IN LIGHT YELLOW.

DS	*Method*	*Strategy*	*Accuracy*	*F1 Macro*	Δ^{-}_{max}	Δ^{-}_{avg-10}	Δ^{-}_{avg-20}	Δ^{-}_{avg-50}	$\lvert\Delta_{avg-all}\rvert$
IEMOCAP	original	-	63.80±0.24	52.44±0.22	-44.79±0.79	-44.41±0.75	-43.68±0.63	-43.01±0.59	2.15±0.04
	w/ random	acquisition	65.91±0.32	53.15±0.35	-42.38±0.93	-42.17±0.89	-41.61±0.77	-39.56±0.74	2.01±0.16
	w/ KNN	acquisition	66.17±0.19	53.59±0.14	-39.85±0.43	-39.80±0.42	-39.02±0.38	-37.19±0.29	1.84±0.03
	w/ clustering	acquisition	65.79±0.48	53.03±0.46	-39.04±0.73	-38.77±0.70	-38.13±0.66	-34.19±0.57	1.39±0.06
	w/ supervision	acquisition	68.19±0.26	55.44±0.27	-40.82±0.39	-40.70±0.37	-40.24±0.24	-38.97±0.19	1.75±0.04
	ours	acquisition	68.45±0.22	55.89±0.21	-33.71±0.29	-33.59±0.28	-33.01±0.21	-29.86±0.15	0.93±0.02
	w/ random	target data++	66.04±1.03	53.67±0.97	-41.13±1.15	-41.04±1.07	-40.55±0.89	-38.98±0.84	1.84±0.55
	w/ KNN	target data++	66.15±0.18	53.64±0.16	-39.72±0.45	-39.51±0.39	-38.79±0.35	-36.35±0.26	1.76±0.04
	w/ clustering	target data++	67.44±0.37	56.17±0.38	-36.19±0.58	-36.03±0.53	-35.28±0.41	-32.03±0.37	0.83±0.05
	ours	target data++	**68.93±0.19**	56.41±0.16	-33.04±0.17	-32.71±0.17	-31.88±0.14	-28.93±0.11	0.59±0.03
	w/ random	regularization	67.51±0.98	55.13±0.95	-40.02±1.01	-39.78±0.96	-39.11±0.82	-37.62±0.69	1.38±0.27
	w/ KNN	regularization	68.03±0.12	55.82±0.15	-38.09±0.34	-37.95±0.31	-37.03±0.25	-35.44±0.19	1.02±0.02
	w/ clustering	regularization	68.39±0.28	56.88±0.25	-35.41±0.47	-35.07±0.43	-34.15±0.39	-31.29±0.30	0.45±0.03
	ours	regularization	68.89±0.15	**56.95±0.13**	**-32.19±0.12**	**-31.04±0.10**	**-29.57±0.09**	**-27.11±0.07**	**0.21±0.02**
	original	all data	67.15±0.13	56.13±0.17	-41.10±0.24	-40.56±0.21	-40.08±0.20	-37.11±0.14	0.88±0.02

DS	*Method*	*Strategy*	*WER*	*CER*	Δ^{-}_{max}	Δ^{-}_{avg-10}	Δ^{-}_{avg-20}	Δ^{-}_{avg-50}	$\lvert\Delta_{avg-all}\rvert$
LIBRISPEECH	original	-	8.05±0.05	2.80±0.04	26.11±0.98	26.02±0.95	25.57±0.89	23.11±0.76	0.29±0.06
	w/ random	acquisition	7.14±0.09	2.38±0.08	17.74±0.61	17.50±0.57	17.12±0.51	16.09±0.44	0.22±0.09
	w/ KNN	acquisition	7.03±0.04	2.32±0.06	14.95±0.47	14.73±0.41	14.19±0.35	13.81±0.32	0.13±0.04
	w/ clustering	acquisition	6.42±0.07	2.01±0.06	12.38±0.52	12.26±0.48	12.07±0.43	11.59±0.37	0.09±0.05
	w/ supervision	acquisition	6.32±0.03	2.01±0.04	17.09±0.58	16.87±0.53	16.22±0.45	14.79±0.36	0.21±0.07
	ours	acquisition	6.31±0.04	1.99±0.04	9.51±0.36	9.38±0.29	9.02±0.25	7.87±0.16	0.07±0.03
	w/ random	target data++	6.89±0.15	2.25±0.14	17.44±0.57	17.28±0.53	17.07±0.42	16.01±0.34	0.17±0.10
	w/ KNN	target data++	6.41±0.07	2.12±0.04	13.19±0.32	13.11±0.26	12.64±0.21	11.38±0.11	0.12±0.06
	w/ clustering	target data++	5.95±0.08	1.92±0.09	11.72±0.38	11.48±0.32	11.09±0.24	10.65±0.20	0.08±0.05
	ours	target data++	5.82±0.04	1.87±0.06	9.27±0.17	9.01±0.14	8.55±0.12	7.72±0.09	0.04±0.02
	w/ random	regularization	6.74±0.17	2.17±0.15	17.51±0.49	17.33±0.47	16.92±0.38	15.84±0.35	0.15±0.09
	w/ KNN	regularization	6.24±0.05	2.05±0.05	13.04±0.26	12.79±0.23	12.08±0.17	10.70±0.12	0.11±0.05
	w/ clustering	regularization	5.80±0.07	**1.83±0.06**	10.98±0.41	10.56±0.38	10.01±0.32	9.47±0.24	0.06±0.03
	ours	regularization	**5.71±0.07**	**1.83±0.05**	**9.12±0.11**	**8.81±0.09**	**8.14±0.08**	**7.59±0.05**	**0.03±0.02**
	original	all data	6.31±0.07	1.98±0.06	14.71±0.85	14.55±0.79	13.98±0.76	13.01±0.68	0.11±0.03

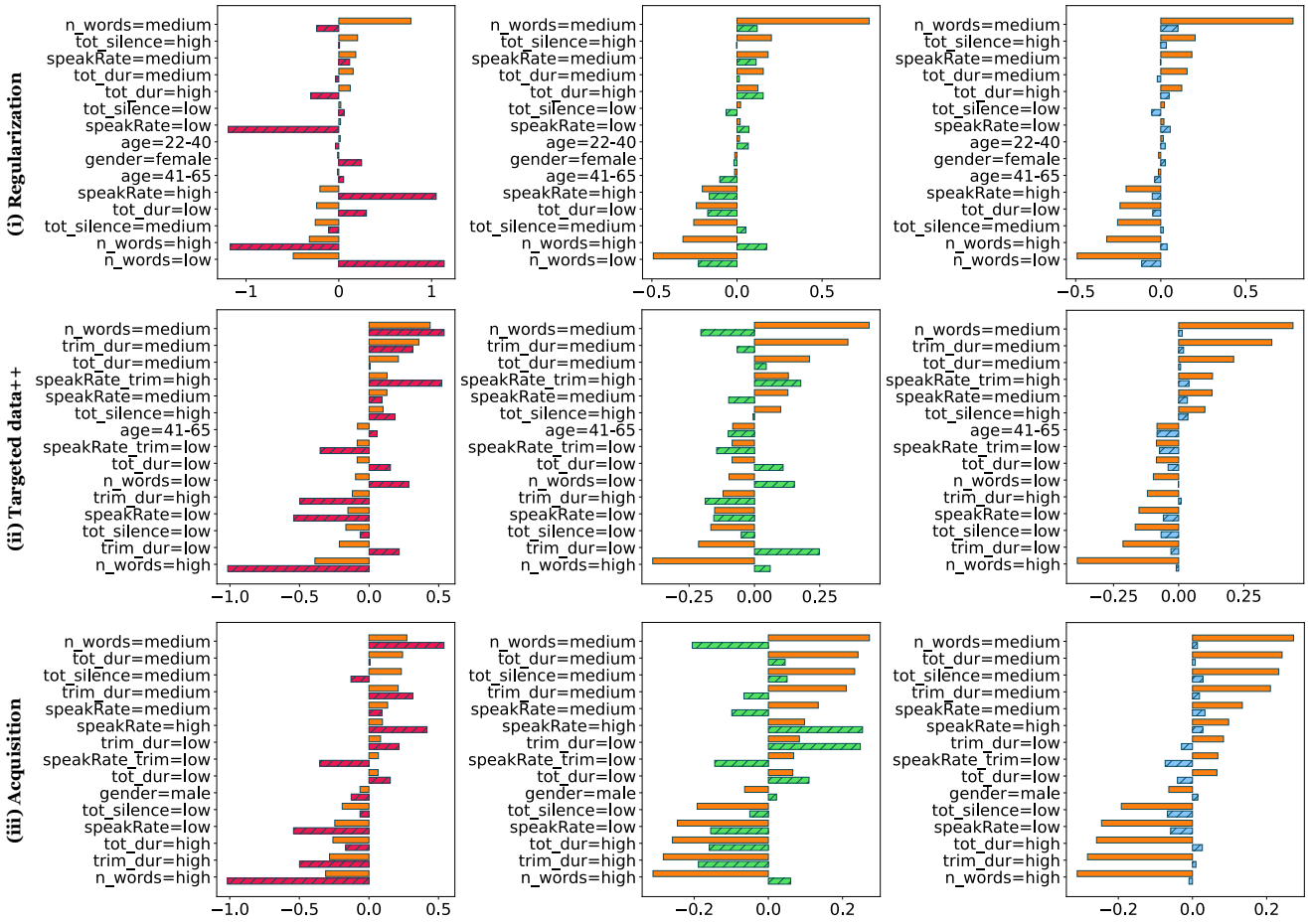
proach, similar to the technique introduced in [37]. Following the model's training phase, we identify instances within the held-out set that the model predicts inaccurately. These instances are labeled as challenging and are subsequently incorporated into the augmented training data. We apply this technique exclusively for post-processing mitigation, as addressing erroneous samples is inherently embedded within standard loss terms during model training. Note that this baseline assumes prior knowledge of the ground truth labels on the held-out set for the tasks at hand.

V. EXPERIMENTAL RESULTS

This section outlines the results and findings of our experiments, focusing on the effectiveness of our mitigation strategies compared to baseline approaches. We assess their performance improvements both overall and in data subgroups (Section V-A). We first evaluate the setup using the hold-out dataset derived from the original training set, allowing us to assess both in- and post-processing methods. We then examine the results when utilizing the entire training set, thus focusing only on the in-processing approach. Finally, we conduct a sensitivity analysis to investigate how varying the number of challenging subgroups impacts post-processing data acquisition and in-processing divergence-aware data augmentation (Section V-B).

A. Mitigation results

1) Comparison against baselines: Tables I and II show the mitigation results of our in- and post-processing strategies compared with the baselines for FSC and ITALIC and for IEMOCAP and LIBRISPEECH, respectively. We use a consistent configuration, using a part of the original train set for actual training and a part of held-out for the data acquisition. This setting ensures the comparability of the results, as we use the same train, validation, and test sets. For each dataset, we report the model's overall and subgroup-based performance results without any mitigation, denoted as 'original.' We then report the results for the in-processing and the two post-processing strategies. For each strategy, we evaluate our approach compared to the baselines, varying how



(a) `w2v2-b` w/ random. (b) `w2v2-b` w/ clustering. (c) `w2v2-b` w/ ours.

Fig. 1. **Global Shapley values (GSV).** Comparison of the top-15 *GSV* of the original model (orange) with (a) random- (shaded red, left), (b) clustering- (shaded green, middle), and (c) *ours* divergence-aware (shaded blue, right) weighting. **Top**: (i) in-processing regularization; **Middle**: (ii) targeted data augmentation; **Bottom**: (iii) post-processing strategy. wav2vec 2.0 base (`w2v2-b`), FSC dataset.

we identify the challenging data samples for the mitigation process. Finally, we outline the results when we train on the complete training set, i.e., when we acquire the entire held-out set, we denote this experiment as 'all data.' In the following, we outline the main outcomes and findings.

Our dual strategy outperforms the baselines. Our in- and post-processing approaches consistently outperform all the baselines, as highlighted in the tables in light yellow. Our approaches not only achieve higher overall performance in accuracy and F1 for IC and ER tasks and WER and CER for the ASR task, but they also significantly improve subgroup-based performance. Specifically, our methods lead to the highest reduction in the divergence of the most underperforming subgroup (Δ^-_{max}), in the average divergence of top 10, 20, and 50 underperforming subgroups (Δ^-_{avg-n}) and of the average absolute divergence across all groups ($\mid \Delta^-_{avg-all} \mid$).

For the post-processing data acquisition, our technique is followed by the error-based baseline for the overall performance. This finding aligns with the intuitive notion that acquiring instances where the model fails can enhance performance. However, clustering emerges as the runner-up method for improving subgroup-based performance. This observation underscores the intuitive strategy of prioritizing the data acquisition efforts towards subgroups (clusters in this case) where the model struggles the most.

For the in-processing techniques, the clustering strategy is the runner-up approach, exhibiting superior results both overall and at the subgroup level compared to the other baselines. This outcome holds for both targeted data augmentation and regularization. Note that we exclude supervised baselines here, as error optimization is intrinsic in the training process and loss function.

In-processing techniques outperform post-processing. Addressing subgroup performance directly during model training proves more effective in mitigating subgroup disparities than operating on an already trained model. As we observe from Tables I and II, the divergence-aware regularization and the targeted data augmentation consistently yield lower scores for Δ^-_{max}, Δ^-_{avg-n}) and $\mid \Delta^-_{avg-all} \mid$ compared to the data acquisition across all evaluated datasets.

Divergence-aware regularization yields the best results. In-processing regularization outperforms the in-process data augmentation and the post-processing technique. The approach demonstrates superior performance both in overall metrics and

TABLE III
MEAN AND STANDARD DEVIATION RESULTS OF THREE RUNS ON THE CONSIDERED IC DATASETS. ORIGINAL FINE-TUNING AND IN-PROCESSING MITIGATION STRATEGIES, INCLUDING REGULARIZATION AND TARGETED DATA AUGMENTATION (TARGET DATA++) WITH K=2, CONSIDERING *all* AVAILABLE TRAINING DATA. BEST RESULTS FOR EACH DATASET ARE IN **BOLD**, SECOND-BEST UNDERLINED; BEST RESULTS FOR EACH DATASET AND STRATEGY IN LIGHT YELLOW.

DS	*Method*	*Strategy*	*Accuracy*	*F1 Macro*	Δ^-_{max}	Δ^-_{avg-10}	Δ^-_{avg-20}	Δ^-_{avg-50}	$\lvert\Delta_{avg-all}\rvert$
FSC	original	-	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
	w/ random	target data++	94.91±0.87	94.46±0.86	-42.62±2.94	-42.51±2.88	-41.80±2.72	-37.19±2.38	0.36±0.24
	w/ KNN	target data++	96.72±0.34	96.15±0.39	-40.01±1.59	-39.59±1.57	-38.61±1.32	-34.09±0.99	0.31±0.08
	w/ clustering	target data++	97.85±0.37	97.59±0.65	-37.57±2.68	-37.21±2.49	-36.13±2.32	-32.75±2.07	0.24±0.11
	ours	target data++	98.46±0.11	98.42±0.17	-27.51±0.56	-27.12±0.52	-26.84±0.48	-22.15±0.43	0.21±0.08
	w/ random	regularization	96.46±0.56	96.29±0.66	-41.31±7.00	-41.31±7.00	-41.14±7.04	-40.66±7.15	0.79±0.94
	w/ KNN	regularization	97.55±0.28	97.38±0.24	-38.29±2.34	-38.02±2.25	-36.56±2.01	-32.15±1.54	0.53±0.06
	w/ clustering	regularization	97.85±0.37	97.59±0.65	-37.57±8.68	-36.28±8.21	-33.69±7.24	-30.74±6.48	0.13±0.02
	ours	regularization	**98.47±0.11**	**98.43±0.14**	**-24.49±0.57**	**-24.49±0.57**	**-24.11±0.51**	**-22.09±0.38**	**0.11±0.01**
ITALIC	original	-	75.71±0.36	73.22±0.33	-47.54±0.79	-47.36±0.76	-46.68±0.47	-41.93±0.00	0.15±0.03
	w/ random	target data++	76.06±0.29	73.36±0.77	-45.82±1.89	-45.34±1.72	-44.65±1.39	-40.82±1.10	0.13±0.09
	w/ KNN	target data++	77.15±0.21	74.03±0.24	-37.87±0.89	-37.12±0.83	-36.41±0.74	-34.04±0.67	0.12±0.04
	w/ clustering	target data++	77.81±0.56	74.19±0.49	-36.73±2.53	-36.19±2.27	-34.15±2.02	-32.58±1.84	0.08±0.02
	ours	target data++	78.01±0.49	74.74±0.35	-30.49±1.77	-30.02±1.52	-27.48±1.47	-24.73±1.21	0.05±0.03
	w/ random	regularization	77.47±0.22	72.76±0.22	-45.11±1.41	-44.99±1.40	-44.24±1.33	-39.58±1.14	0.10±0.01
	w/ KNN	regularization	77.96±0.19	74.12±0.23	-36.39±1.17	-36.14±1.09	-33.87±0.98	-30.05±0.91	0.07±0.02
	w/ clustering	regularization	78.01±0.45	74.45±0.35	-32.81±2.35	-32.73±2.32	-32.13±2.38	-28.97±2.16	0.05±0.03
	ours	regularization	**78.07±0.53**	**74.85±0.30**	**-30.10±1.71**	**-29.64±1.70**	**-27.31±1.66**	**-24.09±2.19**	**0.01±0.04**

at the subgroup level.

2) Analysis of subgroup mitigation process: We further analyze the impact of the mitigation process on subgroup divergence. We can be interested in studying which metadata are generally associated with lower performance (or higher) than the average in the original model and how the mitigation process impacts such divergent behavior. For this analysis, we use the notion of Global Shapley value (GSV), as described in [9]. The GSV estimates the contribution of each metadata (e.g., 'gender=Female') to the divergence across all extracted subgroups. The higher the value, the more the metadata value is associated with different performance than overall ones. For instance, consider the in-processing regularization strategy and the FSC dataset. Figure 1 (top) shows the top-15 metadata values Global Shapley values before (orange) and after mitigation (shaded) for using the (a) random baseline, (b) clustering baseline, and (c) our proposed approach. The random baseline fails to reduce the Global Shapley values, with some values even increasing. This confirms the inability of the random-based mitigation to address subgroup disparities. On the other hand, the clustering-based approach generally reduces the values, showing the benefit of addressing mitigation at the subgroup level. Notably, our approach achieves the most substantial reduction in these global contributions. Similar considerations also apply to the in-processing targeted augmentation strategy, as shown in Figure 1(middle), as well as the post-processing acquisition scenario depicted in Figure 1(bottom). Across all scenarios, our approach consistently demonstrates significantly better performance in minimizing global contributions, effectively flattening them towards zero.

3) In-processing mitigation with the complete training set: Table III shows the mitigation results of our in-processing strategies compared with the baselines using the *complete* training set for FSC and ITALIC. Note that we only consider in-processing techniques as post-processing data acquisition requires separate and unseen data. The results confirm that our approaches overcome the baselines in both overall and subgroup metrics and that regularization is more effective than targeted data augmentation. They also demonstrate that as the number of training samples increases, our in-processing methods enable us to achieve better performance compared to the results shown in Table I, as one might have expected intuitively.

B. Sensitivity analysis

We study how varying the number of challenging subgroups K impacts the data augmentation and data acquisition approaches. We do not examine the effect on the regularization as it does not depend on the number of challenging groups. Figure 2 shows the impact of K on mitigation results for the FSC dataset, with the top part highlighting the post-processing acquisition strategy, and the bottom one the in-processing targeted data acquisition scenario. Specifically, we study the overall performance (F1 Macro) and subgroup performance in terms of average divergence of the top-10 subgroup with lower performance than the average (Δ^-_{avg-10}) and the average absolute divergence across all subgroups ($|\Delta_{avg-all}|$). Considering our approach, lower values of K correspond to higher overall performance. Intuitively, we let the model prioritize addressing the subgroups it struggles with most, resulting in the highest performance improvement. Similarly, lower values of K reduce the average subgroup divergence Δ^-_{avg-10} as we again let the model focus more on a few challenging subgroups. Conversely, as we increase K, we decrease the $|\Delta_{avg-all}|$. This outcome indicates that by targeting a broader

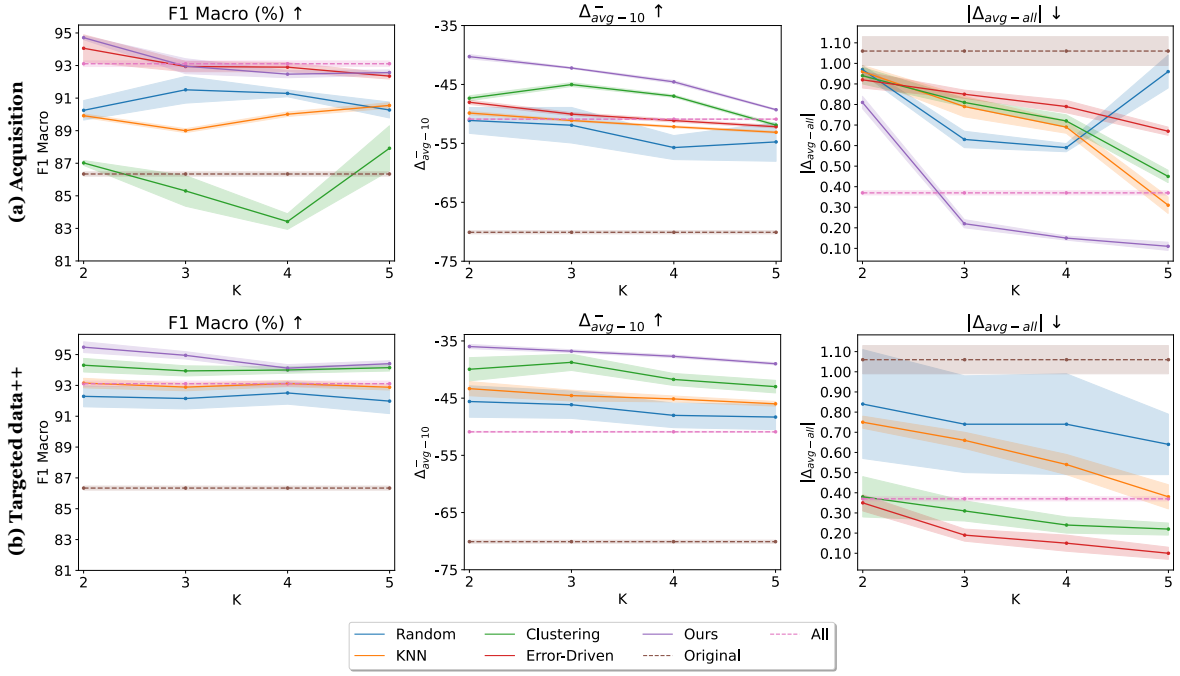


Fig. 2. **Sensitivity Analysis on K**. F1 Macro (left), Δ^{-}_{avg-10} (middle), and $|\Delta_{avg-all}|$ (right) for the considered approaches in (a) the post-processing acquisition (up) and (b) in-processing targeted data augmentation (down) settings, varying K from 2 to 5. wav2vec 2.0 base model, FSC dataset.

range of subgroups for mitigation, the model can address a wider range of subgroup behavior.

The divergence-aware regularization does not depend on the parameter K. Hence, not only allows to achieve the best results generally, but it does not need this parameter setting. This set this regularization as a more suitable and suggested in-processing technique than divergence-aware data augmentation.

VI. Conclusions

This study addresses the critical aspect of mitigating disparities in performance across different population subgroups by proposing a divergence-aware dual mitigation strategy. Our approach automatically identifies subgroups showing a worse performance compared to the overall model behavior and addresses such disparate treatment. We propose both a post-processing method and two in-processing approaches, thus offering versatility and adaptability to diverse real-world scenarios. The post-processing technique mitigates biases of an already trained model by acquiring data samples from underperforming subgroups thanks to their interpretable representation. The in-processing methods address biases during the training itself, and we proposed both targeted data augmentation and divergence-aware regularization. Our experimental results show the efficacy of post-processing targeted sample acquisition in enhancing subgroups and overall model performance of trained models compared to existing baselines. Notably, the in-processing methods show the best results in reducing disparities, with regularization slightly outperforming subgroup-based data augmentation. Our paper offers a comprehensive framework for addressing subgroup disparities at two critical stages of model development, during training and post-training adjustment, offering practitioners versatile tools to mitigate speech model biases.

References

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

[2] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proc. of the National Academy of Sciences*, 2020.

[3] Z. Mengesha, C. Heldreth, M. Lahav, J. Sublewski, and 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*, p. 169, 2021.

[4] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer Speech & Language*, vol. 84, p. 101567, 2024.

[5] J. P. Bajorek, "Voice recognition still has significant race and gender biases," *Harvard Business Review*, vol. 10, 2019.

[6] C. Liu, M. Picheny, L. Sarı, P. Chitkara, A. Xiao, X. Zhang, M. Chou, A. Alvarado, C. Hazirbas, and 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.

[7] Z. Liu, I.-E. Veliche, and F. Peng, "Model-based approach for measuring the fairness in asr," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6532–6536.

[8] O. Niebuhr and A. Michaud, "Speech data acquisition: the underestimated challenge," *KALIPHO-Kieler Arbeiten zur Linguistik und Phonetik*, vol. 3, pp. 1–42, 2015.

[9] A. Koudounas, E. Pastor, G. Attanasio, V. Mazzia, M. Giollo, T. Gueudre, L. Cagliero, L. de Alfaro, E. Baralis, and 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.

[10] E. Pastor, L. de Alfaro, and E. Baralis, "Looking for trouble: Analyzing classifier behavior via pattern divergence," in *Proceedings of the 2021 International Conference on Management of Data*, ser. SIGMOD '21. ACM, 2021, p. 1400–1412.

[11] V. Panayotov, G. Chen, D. Povey, and 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.

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

[13] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and 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. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-2396

[14] A. Koudounas, M. La Quatra, L. Vaiani, L. Colomba, G. Attanasio, E. Pastor, L. Cagliero, and E. Baralis, "ITALIC: An Italian Intent Classification Dataset," in *Proc. INTERSPEECH 2023*, 2023, pp. 2153–2157.

[15] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[16] A. Babu and et al., "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Proc. Interspeech 2022*, 2022.

[17] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[18] A. Koudounas, E. Pastor, G. Attanasio, Luca, L. de Alfaro, and 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.

[19] J. L. Martin and K. Tang, "Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be"," in *Proc. Interspeech 2020*, 2020, pp. 626–630.

[20] R. Tatman and C. Kasten, "Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions." in *Interspeech*, 2017, pp. 934–938.

[21] M. Garnerin, S. Rossato, and L. Besacier, "Gender representation in french broadcast corpora and its impact on asr performance," in *Proceedings of the 1st international workshop on AI for smart TV content production, access and delivery*, 2019, pp. 3–9.

[22] ——, "Investigating the impact of gender representation in asr training data: A case study on librispeech," in *3rd Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 2021.

[23] Z. Liu, I.-E. Veliche, and F. Peng, "Model-based approach for measuring the fairness in asr," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6532–6536.

[24] J. Meyer, L. Rauchenstein, J. D. Eisenberg, and N. Howell, "Artie bias corpus: An open dataset for detecting demographic bias in speech applications," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 6462–6468.

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

[26] L. Sarı, M. Hasegawa-Johnson, and C. D. Yoo, "Counterfactually fair automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3515–3525, 2021.

[27] I.-E. Veliche and 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.

[28] H. Shen, Y. Yang, G. Sun, R. Langman, E. Han, J. Droppo, and A. Stolcke, "Improving fairness in speaker verification via group-adapted fusion network," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7077–7081.

[29] I. Chen, F. D. Johansson, and D. Sontag, "Why is my classifier discriminatory?" *Advances in neural information processing systems*, vol. 31, 2018.

[30] K. H. Tae and S. E. Whang, "Slice tuner: A selective data acquisition framework for accurate and fair machine learning models," in *Proceedings of the 2021 International Conference on Management of Data*, 2021, pp. 1771–1783.

[31] A. Asudeh, Z. Jin, and H. Jagadish, "Assessing and remedying coverage for a given dataset," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 2019, pp. 554–565.

[32] E. Pastor, A. Gavgavian, E. Baralis, and L. de Alfaro, "How divergent is your data?" *Proc. VLDB Endow.*, vol. 14, no. 12, p. 2835–2838, jul 2021. [Online]. Available: https://doi.org/10.14778/3476311.3476357

[33] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and 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.

[34] *SUPERB: Speech Processing Universal PERformance Benchmark*, 2021.

[35] A. Koudounas, E. Pastor, G. Attanasio, V. Mazzia, M. Giollo, T. Gueudre, E. Reale, L. Cagliero, S. Cumani, L. de Alfaro, E. Baralis, and D. Amberti, "Towards comprehensive subgroup performance analysis in speech models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1468–1480, 2024.

[36] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *EMNLP: System Demonstrations*, Oct. 2020.

[37] R. Magar and A. B. Farimani, "Learning from mistakes: Sampling strategies to efficiently train machine learning models for material property prediction," *Computational Materials Science*, vol. 224, 2023.