A Contrastive Learning Approach to Mitigate Bias in Speech Models

Interspeech 2024
1-5 September 2024, Kos, Greece

A Contrastive Learning Approach to Mitigate Bias in Speech Models

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

[†]Politecnico di Torino, Turin, Italy
{firstname.lastname}@polito.it

Abstract

Speech models may be affected by performance imbalance in different population subgroups, raising concerns about fair treatment across these groups. Prior attempts to mitigate unfairness either focus on user-defined subgroups, potentially overlooking other affected subgroups, or do not explicitly improve the internal representation at the subgroup level. This paper proposes the first adoption of contrastive learning to mitigate speech model bias in underperforming subgroups. We employ a three-level learning technique that guides the model in focusing on different scopes for the contrastive loss, i.e., task, subgroup, and the errors within subgroups. The experiments on two spoken language understanding datasets and two languages demonstrate that our approach improves internal subgroup representations, thus reducing model bias and enhancing performance.

Index Terms: model bias, contrastive learning, spoken language understanding, bias mitigation

1. Introduction

Ensuring balanced performance across diverse subgroups of data is a critical aspect of developing fair and unbiased speech models. For instance, a model should perform similarly for all speakers, regardless of factors such as demographics and recording conditions. However, a growing body of work revealed speech models behave differently for different subpopulations [1–11].

This work proposes a novel approach to mitigate performance disparities in data subgroups by directly acting on the latent space representation of samples. To this end, we introduce CLUES, a Contrastive Learning framework for mitigating model biases towards UnderpErforming Subgroups. Specifically, we leverage Contrastive Learning (CL), which has emerged as a significant advancement in representation learning [12, 13]. The idea of CL is to learn representations that place similar samples close together and dissimilar ones further apart. We use this notion to guide the model in learning how to represent samples from the same subgroup close to each other. The intuition is that refining the model representations at the subgroup level enables it to better capture their distinct characteristic, thus mitigating performance disparities.

CLUES employs a three-level contrastive learning loss, with (i) a first contrastive term that operates at the task level, grouping together samples sharing the same class and separating different classes; (ii) a second term to map samples of the same subgroup close together in the embedding space while separating different subgroups; (iii) a third loss term that operates within each subgroup, aggregating correctly predicted samples while setting apart incorrect ones. These losses guide the model in learning representations that capture different scopes, i.e., tasks, subgroups, and errors within subgroups, resulting in more informative embeddings. While the first loss targets the overall model performance, the second and third losses aim to decrease performance disparities between subgroups, thus further enhancing model behavior.

Recent approaches for mitigation propose targeted data augmentation [10], acquiring focused data [8,14], incorporating information from automatically identified subgroups at training time [15], or employing ad-hoc loss functions [5]. However, these methods do not explicitly introduce mitigation by improving the latent representations.

Fairer representations via CL have been beneficial to address disparities for tasks such as text [16, 17] and image classification [18], machine translation [19], and in pre-trained language models on downstream tasks [20]. However, these works address modalities different from speech. Moreover, most of these works focus on improving representations for known protected subgroups. However, disparities may occur for subpopulations that are unknown *a priori*. In response, we mitigate disparities for automatically identified subgroups.

While multiple works adopt CL to train speech models [21–24], the adoption of contrastive learning for fairness in the speech domain remains largely unexplored. To the best of our knowledge, our work is the first to integrate contrastive loss terms to obtain improved representations in speech models.

We experimentally evaluate our approach on two public datasets for intent classification, FSC [25] in English and ITALIC [26] in Italian, with the state-of-the-art transformer models wav2vec 2.0 [27] and XLS-R [28]. The experimental results demonstrate that models trained with our contrastive learning schema develop richer representations that effectively reduce subgroup performance disparities and improve overall performance. Specifically, we reduce the disparity in performance of the most underperforming subgroup by 66.9% (15.5%) for FSC (ITALIC) w.r.t. the baseline models. We also observe an increase of 6.1% (4.8%) in the overall F1 Macro.

2. Methodology

We propose CLUES, a Contrastive Learning framework for mitigating biases toward UnderpErforming Subgroups. CLUES requires defining the subpopulations of interest to guide representation learning. Although CLUES is agnostic to the method adopted for defining subgroups, we discuss two possible approaches in §2.1. Based on the extracted subgroups, we present the three-step contrastive learning schema in §2.2.

10.21437/Interspeech.2024-1219

2.1. Subgroups identification

In this work, we consider two possible subgroup extraction techniques: K-Means [29] clustering and DivExplorer [30].

K-means clustering. K-Means is a commonly adopted clustering algorithm, which assigns one of K clusters to any sample. We consider each cluster as a subgroup. We apply the clustering algorithm to the latent representations of the input points, as extracted by the backbone model used. The cluster extraction is done at the beginning of each training epoch. In this way, the subgroups reflect the evolution of the latent representations throughout the training.

DivExplorer. DivExplorer [30] leverages (typically interpretable) metadata to construct subgroups that satisfy a frequency threshold within the dataset. In the case of speech, these metadata may relate to speaker traits (e.g., gender, age, accent), recording conditions (e.g., utterance duration, noise levels, speaking rate), and task characteristics (e.g., the action, object, and location of the intent). DivExplorer introduces the concept of divergence as the difference in performance between a subgroup and the overall data. We use this notion to assign each point to the most divergent subgroup to which the sample belongs. In this way, we guide CLUES toward mainly improving representations of underperforming subgroups, i.e., the subgroups that diverge the most from the average model behavior.

2.2. CLUES

Our approach aims to *hint* the model toward learning embeddings that simultaneously achieve: (i) the separation of samples according to a classification task (e.g., intent classification), (ii) the separation of samples according to the identified subgroups, and (iii) an equitable representation of over- and underperforming subgroups within the classification task. To this end, CLUES consists of three separate contrastive learning levels: task, subgroup, and error. At each level, we employ the multi-similarity (MS) loss [13] to selectively contrast sample pairs based on their affinity. This loss has already proven to be effective in speech tasks [31]. For any sample (referred to as anchor in CL), we identify positive and negative samples based on the various levels' criteria. The contrastive loss works toward getting positive samples closer to the anchor and negative samples further away from it. For each level, the corresponding loss is computed as the sum of positive and negative terms:

$$\mathcal{L}_{MS} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{\alpha}\log[1+\sum_{p\in\mathcal{P}_i}e^{-\alpha(S_{ip}-\lambda)}] + \frac{1}{\beta}\log[1+\sum_{n\in\mathcal{N}_i}e^{\beta(S_{in}-\lambda)}] \quad (1)$$

Where m is the batch size, $\mathcal{P}_i$ and $\mathcal{N}_i$ denote the sets of positive and negative samples for the anchor i, S_{ip} and S_{in} are the similarities between i and its positive/negative pairs, and α, β, λ control pair weighting. The three loss terms introduced, which are summarized in Figure 1, are described as follows.

Task-level Contrastive Learning. We introduce a first MS contrastive loss function focused on the classification task level $\mathcal{L}_t$. This loss groups samples sharing the same class and separates samples belonging to different classes, and aims at improving the separability of the samples in the downstream task.

Subgroup-level Contrastive Learning. We use a second MS contrastive loss function $\mathcal{L}_s$ to guide the learning of subgroup-level representations. We leverage the subgroups identified within the data, as described in Section §2.1. In particular, we choose as positive pairs points belonging to the same subgroup, and as negative pairs points belonging to different ones. This encourages an internal representation that is aware of the identified subpopulations, not only of the task being addressed.

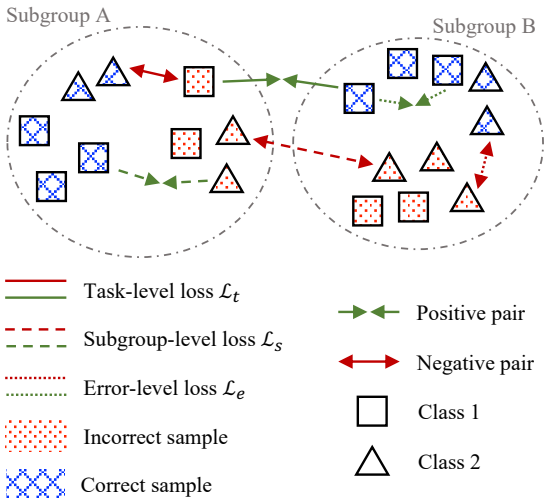

Figure 1: *Summary of the action of the three contrastive loss terms on a toy example, comprised of 2 subgroups (A, B) and a binary classification task (square/triangle).*

Error-level Contrastive Learning. Finally, we introduce a contrastive loss term $\mathcal{L}_e$ that considers intra-subgroup errors on the classification task. The outcome of a model for each sample can be represented as a binary (correct/incorrect) result. Within each subgroup, we define as positive the pairs of samples that have obtained the same outcome, and as negative the ones with different outcomes. This introduces a bipartition within each subgroup: one of the partitions contains all correctly predicted samples, and the other contains all incorrectly predicted ones. As samples get predicted correctly, the $\mathcal{L}_e$ loss moves them toward the correctly predicted partition.

Final Loss. We define our overall training objective as the aggregation of these multi-similarity losses, along with a conventional classification loss $\mathcal{L}_{cls}$ (e.g., cross-entropy):

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_t\mathcal{L}_t + \lambda_s\mathcal{L}_s + \lambda_e\mathcal{L}_e \quad (2)$$

where λ_t, λ_s and λ_e are coefficients that regulate the relative importance of each loss term. We identify the best values for these coefficients with a tuning phase using a validation set.

3. Experimental Setup

In this section, we detail the setup used for the experiments[1].

Datasets. We evaluated our approach on two public intent classification datasets, FSC [25] for English and ITALIC [26] for Italian. FSC includes 30,043 utterances annotated with action, object, and location defining intents, while ITALIC contains 16,521 samples with action and scenario denoting intents. Both datasets split speakers across train, validation, and test sets.

Metadata. We considered demographic, speaking and recording conditions, and intent-related metadata, following the

[1] https://github.com/koudounasalkis/CLUES

Table 1: *Mean ± std of three runs on* FSC *and* ITALIC. *Comparison of original fine-tuning, data augmentation [10], adversarial loss [5], data acquisition [14], and CLUES. For all metrics, higher is better. The best results are in* ***bold****, the second-best are* <u>*underlined*</u>*.*

DS	*Approach*	*Subgroups*	*Accuracy*	*F1 Macro*	Δ^-_{max}	S	$S^{\pm}$
FSC	w2v2-b original	-	93.419±0.169	93.110±0.168	-53.179±0.147	0.737±0.049	0.318±0.084
	w/ data++ [10]	-	94.909±0.870	94.460±0.861	-42.623±2.939	0.757±0.031	0.309±0.021
	w/ adversarial [5]	K-Means	<u>98.591±0.210</u>	98.507±0.187	-26.141±0.117	0.819±0.016	0.401±0.017
	w/ adversarial [5]	DivExplorer	98.486±0.105	98.311±0.109	-24.512±0.145	0.814±0.015	0.389±0.021
	w/ data acquisition [14]	K-Means	96.511±0.309	95.983±0.338	-32.488±0.461	0.749±0.048	0.324±0.053
	w/ data acquisition [14]	DivExplorer	96.719±0.215	96.054±0.274	-22.692±0.316	0.755±0.029	0.334±0.032
	w/ CLUES	K-Means	98.567±0.145	<u>98.514±0.141</u>	<u>-21.410±0.393</u>	<u>0.848±0.019</u>	<u>0.516±0.015</u>
	w/ CLUES	DivExplorer	**98.789±0.104**	**98.761±0.095**	**-17.581±0.433**	**0.894±0.012**	**0.525±0.014**
ITALIC	XLSR-300 original	-	75.711±0.360	73.218±0.329	-47.541±0.789	0.319±0.064	-0.221±0.081
	w/ data++ [10]	-	76.062±0.289	73.361±0.771	-45.820±1.892	0.323±0.099	-0.213±0.091
	w/ adversarial [5]	K-Means	77.499±0.315	75.014±0.437	-44.117±0.654	0.447±0.098	-0.104±0.089
	w/ adversarial [5]	DivExplorer	77.201±0.641	74.840±0.527	-42.535±0.714	0.461±0.086	-0.094±0.075
	w/ data acquisition [14]	K-Means	76.308±0.512	74.016±0.505	-41.918±0.672	0.375±0.024	-0.210±0.027
	w/ data acquisition [14]	DivExplorer	77.510±0.441	75.201±0.384	<u>-41.005±0.510</u>	0.389±0.019	-0.197±0.022
	w/ CLUES	K-Means	**80.561±0.554**	<u>76.104±0.317</u>	-43.010±0.892	<u>0.512±0.034</u>	<u>0.214±0.028</u>
	w/ CLUES	DivExplorer	<u>79.230±0.810</u>	**76.721±0.201**	**-40.150±0.963**	**0.539±0.025**	**0.241±0.023**

Table 2: *Ablation study on $\mathcal{L}_t$, $\mathcal{L}_s$, and $\mathcal{L}_e$. The best results are in* ***bold****. Subgroups extracted with DivExplorer.*

DS	*Approach*	*F1 Macro*	Δ^-_{max}	S	$S^{\pm}$
FSC	w2v2-b	93.110	-53.179	0.737	0.318
	w/ $\mathcal{L}_t$	98.105	-48.714	0.759	0.309
	w/ $\mathcal{L}_s$	98.434	-26.551	0.848	0.419
	w/ $\mathcal{L}_t + \mathcal{L}_s$	98.430	-33.124	0.814	0.371
	w/ $\mathcal{L}_s + \mathcal{L}_e$	98.452	-19.112	0.861	0.501
	w/ $\mathcal{L}_s + \mathcal{L}^*_e$	98.106	-20.014	0.865	0.487
	w/ CLUES	**98.761**	**-17.581**	**0.894**	**0.525**
ITALIC	XLSR-300	73.218	-47.541	0.319	-0.221
	w/ $\mathcal{L}_t$	76.075	-49.540	0.346	-0.222
	w/ $\mathcal{L}_s$	76.326	-45.391	0.452	-0.196
	w/ $\mathcal{L}_t + \mathcal{L}_s$	76.279	-47.443	0.439	-0.208
	w/ $\mathcal{L}_s + \mathcal{L}_e$	76.620	-43.288	0.490	0.215
	w/ $\mathcal{L}_s + \mathcal{L}^*_e$	76.319	-43.329	0.499	0.123
	w/ CLUES	**76.721**	**-40.150**	**0.539**	**0.241**

metadata-enrichment proposed in [9]. When using DivExplorer, we explored all subgroups with a minimum frequency of 0.03, while for K-Means we considered K=10 for ITALIC and K=20 for FSC. These configurations have been found to achieve the best performance on the target datasets [14].

Models. We fine-tuned the pre-trained wav2vec 2.0 [27] base and multilingual XLS-R [28] models on the FSC and ITALIC datasets, respectively. The pre-trained checkpoints were obtained from the Hugging Face hub [32] and served as our baselines. We followed fine-tuning procedures from relevant literature [33]. We adhere to standard procedures in CL, selecting positive and negative sample pairs within each batch to optimize model performance. Further details about models, hyperparameters, and fine-tuning are available in the project repository.

Metrics. We assessed the overall model performance with accuracy and macro F1 score. We also considered the highest negative subgroup divergence (Δ^-_{max}), where the divergence Δ of a subgroup is the difference between that subgroup accuracy and the overall one. Δ^-_{max} evaluates how well the model can reduce differences in performance between subgroups and thus mitigate bias. Finally, we evaluated the quality of the latent space representation in terms of Silhouette [34], which quantifies intra-subgroup cohesion and inter-subgroup separation. We measured both the Silhouette w.r.t. the adopted subgroups (S), and the Silhouette w.r.t. the partitions of correctly/incorrectly predicted samples within the subgroups ($S^{\pm}$).

Baselines. We evaluated CLUES against several competitors. We explored a standard data augmentation scenario [10]. Additionally, following [5], we experimented with an extra adversarial loss that aims to predict whether an utterance belongs to an underperforming subgroup or not[2]. Finally, we included a baseline inspired by the work of [14], simulating a scenario where additional data from the same distribution as the existing dataset is available. In this baseline, samples from underperforming subgroups are selectively added. For [5], [14], and CLUES, the subgroups are identified using both K-means and DivExplorer.

4. Experimental Results

We report the main experimental results in Table 1. The outcomes obtained are consistent across the two datasets. Both K-Means and DivExplorer enabled CLUES to perform well in general, demonstrating its ability to deal with different subgroup extraction techniques. Since DivExplorer-based results are generally better, we focus the discussion on this approach. The results show that CLUES achieved the best overall model performance in terms of both accuracy and F1 score. On top of improving the overall performance, we also observed a decrease in maximum negative divergence. For instance, on FSC, the most underperforming subgroup deviates by 53.2 accuracy points from the overall performance. CLUES brings the divergence of the worst-performing subgroup down to -17.6% instead. This decrease also surpasses the one obtained with the acquisition of new, targeted data. We mainly attribute this improvement to $\mathcal{L}_s$ and $\mathcal{L}_e$, which focus on reducing the intra-subgroup dispersion. We used the Silhouette to measure the

[2]While the authors of [5] employ an extra loss for distinguishing between native and non-native speakers, we propose to discern utterances belonging to underperforming and non-underperforming subgroups.

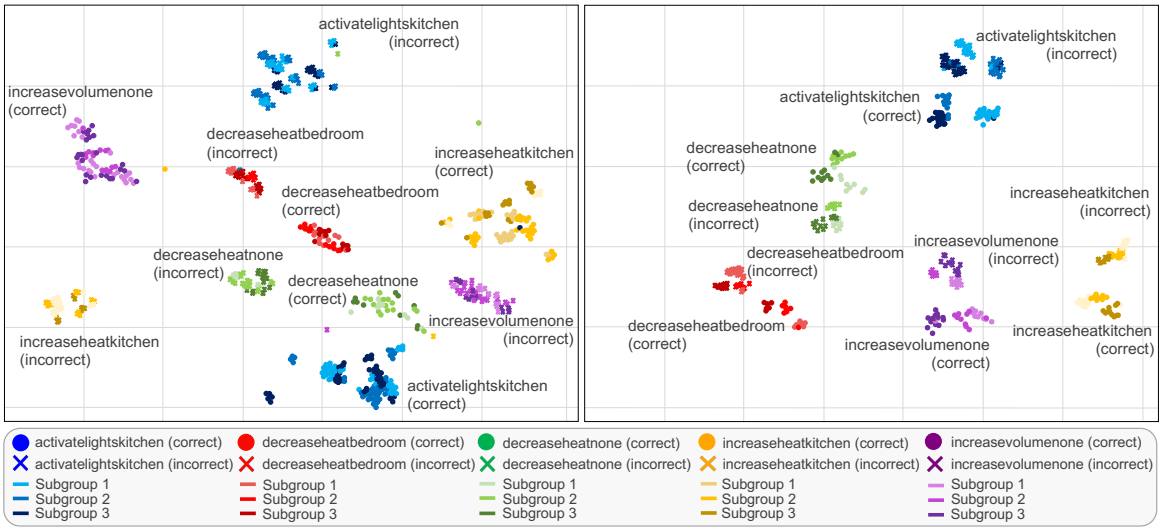

Figure 2: FSC. *5 most-frequent intents (main color) and 3 most frequent subgroups (shades of the same color). t-SNE visualization of the original model (left) and CLUES (right). Correct samples reported as circles, incorrect ones as crosses. Best viewed in color.*

quality of the latent space representations. Also in these terms, CLUES achieved by far the best performance for both S and $S^{\pm}$, thanks to the $\mathcal{L}_s$ (for S) and $\mathcal{L}_s + \mathcal{L}_e$ terms (for $S^{\pm}$).

Ablation Study. We conducted an ablation study to quantitatively assess the impact of each proposed loss term. Table 2 summarizes these results. All loss terms produce an improvement in accuracy and F1 score over the original model. The best improvement occurs when all loss terms are combined. The reduction in divergence (Δ^{-}_{max}) mainly occurs when the $\mathcal{L}_s$ term is introduced. Indeed, this term acts by grouping together points belonging to the same subgroup, thus producing more easily separable latent representations. We observed an additional decrease in divergence and an improvement in terms of overall performance when adding $\mathcal{L}_e$. For completeness, we additionally explored the option where the error-level loss merges correct and incorrect points together (instead of separating them). We refer to this alternative loss as $\mathcal{L}^*_e$. The results with $\mathcal{L}^*_e$ are still satisfactory, indicating that the intra-subgroup action is still effective. However, $\mathcal{L}_e$ has a better effect on performance w.r.t. $\mathcal{L}^*_e$. Finally, as expected, we note how the introduction of the error-level produced the largest improvement in terms of $S^{\pm}$.

Qualitative Analysis. We provide a qualitative analysis of the effect of CLUES on latent representations. Figure 2 presents a t-SNE [35] visualization that compares the original (wav2vec 2.0) vector space against the CLUES-tuned approach. We use the FSC dataset and select the 5 most frequent intents. For each intent, we visualize the samples belonging to the three most frequent subgroups. As can be expected, both approaches tend to cluster together points sharing the same target intent. However, the original embedding produces generally less cohesive clusters, since correctly and incorrectly predicted samples belonging to the same intent are not placed close together. Additionally, points belonging to the same subgroup are not necessarily close together in the original space. Instead, the contrastive terms introduced by CLUES produce more consistent groups of samples according to both the intent class and the subgroups. We argue that this increase in group cohesion is the reason behind the observed general improvement in performance. For instance, the samples for the *increase heat kitchen* intent (in yellow) are particularly spread in the original representation – with correct and incorrect samples being far away from one another and with little separation between subgroups. By contrast, the CLUES-tuned version groups all relevant samples together into a more cohesive cluster on 3 levels. On a broader level, all points belonging to the intent are close together (e.g., all yellow points). Within each intent, we can identify two partitions for correctly and incorrectly predicted samples (top and bottom of the yellow cluster). Finally, within each partition, the three subgroups form three cohesive sub-clusters.

5. Discussion

We introduced CLUES, a contrastive learning framework to mitigate biases in underperforming subgroups. Through a multi-level contrastive schema, CLUES steers the model towards a balanced understanding of both the classification task and underperforming subgroup modeling. The results showed that CLUES obtains an improvement in model performance and a reduction in subgroup divergence, outperforming competitors. The ablation study confirmed that the loss terms act as desired on the evaluation metrics. As expected, the combination of all loss terms produced the best results across all metrics. Finally, we verified that the latent representations obtained with CLUES are more consistent with the desired behavior, as shown with quantitative (via Silhouette) and qualitative (via t-SNE) results.

Limitations. In this work, we presented experimental results only on intent classification. However, our methodology could be readily extended to other classification tasks. We additionally plan on extending CLUES to other supervised tasks, such as Automatic Speech Recognition.

6. Acknowledgments

The authors thank Giuseppe Averta and Moreno La Quatra for the useful discussions, and Sara Papi for her valuable support during the paper writing. This work is partially supported by the FAIR - Future Artificial Intelligence Research (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, both 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.

7. References

[1] R. Tatman and C. Kasten, "Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions." in *Proc. INTERSPEECH*, 2017.

[2] J. L. Martin and K. Tang, "Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be"," in *Proc. INTERSPEECH*, 2020.

[3] 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, 2021.

[4] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," *arXiv preprint arXiv:2103.15122*, 2021.

[5] Y. Zhang, Y. Zhang, B. M. Halpern, T. Patel, and O. Scharenborg, "Mitigating bias against non-native accents," in *Proc. INTERSPEECH*, 2022.

[6] Z. Liu, I.-E. Veliche, and F. Peng, "Model-based approach for measuring the fairness in asr," in *ICASSP*, 2022.

[7] 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.

[8] 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.

[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] Y. Zhang, A. Herygers, T. Patel, Z. Yue, and O. Scharenborg, "Exploring data augmentation in bias mitigation against non-native-accented speech," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[11] 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.

[12] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[13] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *CVPR*, 2019.

[14] A. Koudounas, E. Pastor, G. Attanasio, 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.

[15] I.-E. Veliche and P. Fung, "Improving fairness and robustness in end-to-end speech recognition through unsupervised clustering," in *ICASSP*, 2023.

[16] J. Chi, W. Shand, Y. Yu, K.-W. Chang, H. Zhao, and Y. Tian, "Conditional supervised contrastive learning for fair text classification," in *EMNLP 2022*.

[17] A. Shen, X. Han, T. Cohn, T. Baldwin, and L. Frermann, "Contrastive learning for fair representations," *arXiv preprint arXiv:2109.10645*, 2021.

[18] Y. Hong and E. Yang, "Unbiased classification through bias-contrastive and bias-balanced learning," *NeurIPS*, 2021.

[19] M. Lee, H. Koh, K.-i. Lee, D. Zhang, M. Kim, and K. Jung, "Target-agnostic gender-aware contrastive learning for mitigating bias in multilingual machine translation," *arXiv preprint arXiv:2305.14016*, 2023.

[20] X. Dong, Z. Zhu, Z. Wang, M. Teleki, and J. Caverlee, "Co$ˆ2$PT: Mitigating bias in pre-trained language models through counterfactual contrastive prompt tuning," in *EMNLP*, 2023.

[21] H. Al-Tahan and Y. Mohsenzadeh, "Clar: Contrastive learning of auditory representations," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2530–2538.

[22] R. Ye, M. Wang, and L. Li, "Cross-modal contrastive learning for speech translation," in *NAACL HLT*, 2022.

[23] N. Vaessen and D. A. van Leeuwen, "The effect of batch size on contrastive self-supervised speech representation learning," *arXiv preprint arXiv:2402.13723*, 2024.

[24] T. Han, H. Huang, Z. Yang, and W. Han, "Supervised contrastive learning for accented speech recognition," *arXiv preprint arXiv:2107.00921*, 2021.

[25] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," in *Proc. INTERSPEECH*, 2019.

[26] 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.

[27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

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

[29] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[30] 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.

[31] M. La Quatra, A. Koudounas, E. Baralis, and S. M. Siniscalchi, "Speech analysis of language varieties in Italy," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 15 147–15 159.

[32] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, and A. M. et al., "Transformers: State-of-the-art natural language processing," in *EMNLP: System Demonstrations*, Oct. 2020.

[33] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, and K. L. et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. INTERSPEECH*, 2021.

[34] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[35] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.