LEVERAGING CONFIDENCE MODELS FOR IDENTIFYING CHALLENGING DATA SUBGROUPS IN SPEECH MODELS

Alkis Koudounas[♣], *Eliana Pastor*[♣], *Vittorio Mazzia*[♡], *Manuel Giollo*[♡], *Thomas Gueudre*[♡], *Elisa Reale*[♡], *Giuseppe Attanasio*[♢], *Luca Cagliero*[♣], *Sandro Cumani*[♣], *Luca de Alfaro*[♠], *Elena Baralis*[♣], *Daniele Amberti*[♡]

[♣]Politecnico di Torino, Turin, Italy, [♡]AGI, Amazon, Turin, Italy
[♢]Bocconi University, Milan, Italy, [♠]University of California, Santa Cruz, CA, USA

ABSTRACT

State-of-the-art speech models may exhibit suboptimal performance in specific population subgroups. Detecting these challenging subgroups is crucial to enhance model robustness and fairness. Traditional methods for subgroup identification typically rely on demographic information such as age, gender, and origin. However, collecting such sensitive data at deployment time can be impractical or unfeasible due to privacy concerns.

This paper introduces a novel Challenging Subgroup Identification model (CSI) to (i) automatically predict if an utterance belongs to a challenging subgroup and (ii) provide an interpretable representation of this subgroup. CSI exploits confidence models (CMs) to encode information about sources of errors, as CMs assess model certainty of predictions, providing insights into output reliability. CM fine-tuning based on challenging subgroup identification techniques allows accurate subgroup identification. CSI leverages demographic features only during its training, avoiding the need for sensitive data collection at deployment time. Experimental results on the automatic speech recognition and intent classification tasks show CSI effectiveness in identifying challenging subgroups and providing interpretable subgroup descriptions. These findings highlight CSI as a valuable tool for improving the robustness and fairness of speech models in real-world applications.

Index Terms— Confidence models, Challenging subgroups, Divergence, Trustworthy AI

1. INTRODUCTION

State-of-the-art speech models addressing Intent Classification (IC) and Automatic Speech Recognition (ASR) are known to perform suboptimally on specific population subgroups [1, 2, 3, 4, 5]. Detecting the most challenging subgroups for a given speech model is particularly relevant to improve its robustness and fairness [6].

In the context of machine learning, any model output P can be enriched with a confidence score. This score estimates the likelihood that P is correct either by using some model-specific uncertainty estimate [7], or by using some auxiliary confidence model (CM) trained for predicting model expected error rate [8]. CMs are widely used in ASR and Natural Language Processing (NLP) to generate confidence scores both for individual words and complete utterances and sentences, and they were shown to be able to mitigate model performance gaps on cohorts [1].

To identify challenging subgroups, traditional approaches [4, 9] analyze demographic features such as age, gender, and accent and correlate them with model performance. While this approach could be feasible at training time on a selection of utterances, collecting demographic information at testing time is not always practicable due to privacy reasons.

The work recently proposed in [6] is, to the best of our knowledge, the first attempt to disentangle speaker-related and acoustic information in challenging subgroup identification. It applies clustering on utterance-level embeddings precomputed by a speaker ID model trained on a public dataset. However, the embedding representation is not interpretable and the clustering model is potentially sensitive to noise.

To overcome these issues, in this work we design, train, and test a new *Challenging Subgroup Identification* model (CSI, in short) on top of traditional confidence models. CSI assigns to a given test utterance its most likely challenging subgroup (if any). While the challenging subgroups under consideration are generated at training time by a state-of-the-art subgroup identification approach [2] leveraging demographic features, such sensitive information is no longer required at testing time as CSI prediction models already incorporate the underlying knowledge.

We run experiments on two publicly available datasets, namely LIBRISPEECH [10] for ASR and FLUENT SPEECH COMMANDS [11] for IC. The empirical findings demonstrate the effectiveness of our methodology in delineating the most challenging subgroups and highlight performance improvements achieved by incorporating the confidence model[1].

[1]Code repository: `https://github.com/koudounasalkis/Leveraging-CMs-for-Problematic-Subgroups`

2. RELATED WORK

Challenging subgroup identification. The automated identification of challenging population subgroups for speech models has been previously explored using unsupervised clustering [1, 6]. In [6], utterance-level embeddings from a speaker ID model trained on public data are clustered. Since public data includes demographic metadata, the encoder inherently incorporates privacy-sensitive information. However, the clustering approach lacks interpretability and is sensitive to outliers. To address this, we suggest using explainable subgroups instead of embedding clusters.

The work in [2] adopts DIVEXPLORER [12, 13] to extract interpretable subgroups. The concept of divergence enables the identification of subgroups on which the speech model performance is unexpected (positively or negatively). However, [2] assumes prior knowledge of demographic metadata at testing time. To overcome this limitation, we introduce a novel predictive model fine-tuned on a CM. This model effectively recognizes challenging subgroups for new utterances at deployment time without accessing metadata.

Confidence models. Relevant works focused on improving performance by managing errors and reducing word error rates [14, 15, 16, 17]. Prominent strategies include the integration of sequence-level confidence classifiers [15], Heterogeneous Word Confusion Networks [14], confidence estimation modules [17], and self-attention-based models [16]. Few attempts also adopted confidence scores in downstream tasks such as data selection for model adaptation, identification of rare words, and semi-supervised learning [17, 18, 19]. Our approach fine-tunes CMs to address challenging subgroup identification as a downstream task, leveraging their error-based predictive abilities.

3. METHOD

Let M be a model trained for a task, like IC or ASR. Our goal is to detect and explain challenging population subgroups early. At deployment time, when M processes unseen data, we would like to know in advance which subgroups M will more likely misclassify. We also aim to provide end-users with interpretable descriptions of these challenging subgroups. The objective is to develop a system to predict the challenging subgroup to which an utterance belongs (or "none" if not challenging) for model M. We thus introduce a novel *"Challenging Subgroup Identification"* (CSI) model, which is fine-tuned from a Confidence Model (CM). Thanks to this fine-tuning, the model can effectively use the encoded information about sources of errors from the CM.

The pipeline, depicted in Figure 1, has three main steps:

1) *Confidence Model (CM) Training.* We pretrain a CM tailored to the task under analysis. The CM captures valuable information about potential sources of errors in model M.

2) *(Prior) Extraction of Challenging Subgroups.* We extract challenging subgroups for model M using a dataset enriched with relevant metadata, including demographics, speech-related, and dataset- or task-dependent information. We identify challenging subgroups through DIVEXPLORER [13], following the approach presented in [2].

3) *Learning the Challenging Subgroup Identification Model.* We fine-tune the CM-based model to predict, for each utterance, the specific challenging subgroup it belongs to. The identification model gains insights into potential challenges by utilizing encoded information from the CM.

Confidence Model (CM) Training. Given a model M trained for a specific task (e.g., IC), we train a Confidence Model (CM) to predict if the model will make a correct or incorrect prediction for a given sample.

Let X be an input dataset of utterances for the task under analysis. Given the original dataset X, we derive a transformed dataset $\mathcal{Z}$ to train the CM model, composed of input features and error-based target labels. The input features for the CM encode the following components: (i) features related to the level of uncertainty, including the length of the n-best list and output probabilities, (ii) acoustic embeddings representing the (last or average of the) hidden states of the model, and (iii) speech metadata, such as the number of words, number of pauses, and speaking rate. We then construct the error-based target labels by annotating each utterance with a label equal to 1 if model M predicts it correctly and 0 if incorrectly. In the context of ASR, label 1 corresponds to utterances for which the Word Error Rate (WER) metric attains a perfect score of 0.0, while label 0 pertains to all other instances.

We train a Confidence Model on $\mathcal{Z}$, using the standard approach of splitting $\mathcal{Z}$ into train, validation, and test subsets. The training and validation sets are used to set and fine-tune the model parameters. By training the CM on this enriched dataset, we aim to equip it with domain-specific knowledge and model awareness tailored to the particular analysis task.

Extraction of Challenging Subgroups. In this step, we extract the challenging subgroups from the dataset under analysis. The challenging subgroups will be used to train the CSI model in the last step. We adopt the DIVEXPLORER method to extract the challenging subgroups, following [2]. Given a set of interpretable metadata that describes utterances, DIVEXPLORER extracts all subgroups adequately represented in the data and computes their performance difference from the overall one on the entire dataset, denoted as divergence. The selection of adequately represented subgroups, denoted as frequent, is based on a frequency threshold.

Following [2], we first enrich the dataset with metadata to enable extraction. We use demographic, speech, and dataset- and task-specific metadata. We assume we can access demographic information during the training time phase. Including demographic metadata allows us to develop, at Step 3, a model that considers sensitive attributes and predicts sub-

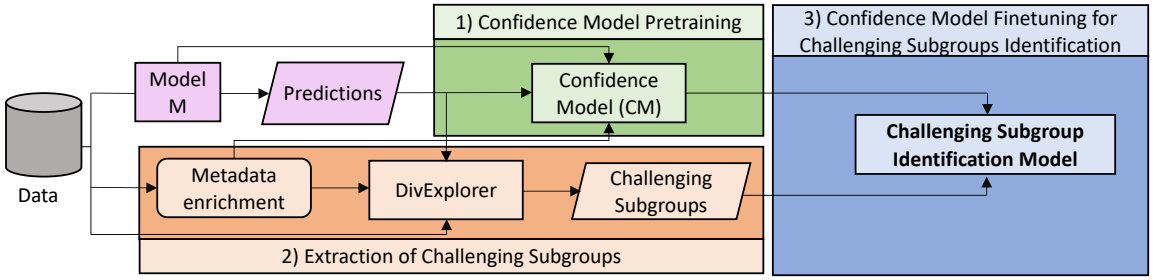

Fig. 1. Pipeline to build the Challenging Subgroup Identification model (CSI).

groups that include these attributes. This is relevant as such sensitive information may be unavailable during inference.
Each extracted subgroup is a conjunction of metadata-value pairs, also denoted as *itemset*. For example, {gender= female, duration>10s} identifies all utterances lasting more than 10 seconds pronounced by women. A divergence in accuracy of -10% indicates that this subgroup has an accuracy lower by 10% than the overall dataset. Among all frequent subgroups obtained by DIVEXPLORER, we focus only on the top K challenging subgroups, i.e., the top K with performance lower than average. To address the issue of redundancy and overlapping subgroups, we adopt the redundancy pruning available in DIVEXPLORER. When two overlapping itemsets have a divergence lower than a specified threshold, we keep the shorter one and remove the redundant duplicate (e.g., between {gender=female, duration>10s} and {gender=female}, we keep the second). This ensures that we retain unique and diverse challenging subgroups.

Challenging Subgroup Identification Model. In the final step, we train the CSI model to predict the challenging subgroup to which each utterance belongs by fine-tuning the CM developed in Step 1.
We label our transformed set $\mathcal{Z}$ with respect to the challenging subgroups. Recall that we derive $\mathcal{Z}$ by transforming the original dataset using the following features: speech metadata, hidden states (acoustic embeddings of the model), and output probabilities. We then label each utterance in $\mathcal{Z}$ with (i) the ID of the challenging subgroup it belongs to, or (ii) the "Non-challenging ID" (0) if the utterance does not belong to any challenging subgroup. An utterance may belong to multiple challenging groups. In this case, we assign the label based on the most divergent subgroup. Our problem is thus remapped to a multi-class classification task.

4. EXPERIMENTAL SETTINGS

Preliminaries. We divide the input data into train, validation, and test sets based on official splits. The train and validation sets are used for model training and tuning in steps 1 to 3. We adopt the test set only for the final evaluation. We empirically analyze CSI outcomes varying subgroups' number (K).

Datasets. We assess our approach on two publicly available datasets: FLUENT SPEECH COMMANDS (FSC) [11] and LIBRISPEECH (LS) [10]. FSC [11] is a widely utilized benchmark for the IC task. The test set includes 3793 audio samples by 10 speakers covering 31 distinct intents. These utterances are characterized by action, object, and location, whose combination delineates the intent. LIBRISPEECH [10] consists of audio recordings for the ASR task. We use the *"clean-360"* version, which includes 360 hours of clean audio samples. The test set includes 2620 samples by 40 speakers. We ensure that each subgroup in the test set comprises at least 100 utterances, and we set the minimum frequency and redundancy thresholds accordingly.

Models. We consider the wav2vec 2.0 [20] base. For FSC, we use the public fine-tuned checkpoints [21], while for LIBRISPEECH, we follow fine-tuning procedures and guidelines from relevant literature [21]. The models achieve 91.72% accuracy on FSC and 6.06% WER on LIBRISPEECH.

CM Training. The architecture of the confidence model includes two hidden layers with GELU activation function, dropout, and normalization layers. We initialize the layers through the Kaiming normal initialization technique. The CM model is trained over a maximum of 10,000 epochs, subject to an early stopping criterion, with NAdam optimizer and a learning rate of 5e-3. We use the Cross-Entropy (CE) loss for FSC. For LIBRISPEECH, we include an additional term, a Mean Squared Error (MSE), using WER as an extra target. The resulting objective function is a weighted combination of CE and MSE losses, described by: $\mathcal{L}_{tot} = \alpha\mathcal{L}_{CE} + (1-\alpha)\mathcal{L}_{MSE}$ where α is set to 0.6.

Metrics. To assess the performance of our approach, we use the AUC and Accuracy metrics for Step 1, and the F1 macro score and the Error Rate (ERR) for Step 3 using the metadata of the test set to derive the ground truth challenging groups.

Table 1. Results of the CM error identification (Step 1) and challenging subgroup identification (Step 3). FSC and LIBRISPEECH (LS) datasets. Best results are highlighted in boldface.

	CM Performance		Challenging Subgroups Identification							
	AUC	Accuracy	Approach	K = 2		K = 3		K = 4		K = 5	
				ERR ↓	**F1** ↑	**ERR** ↓	**F1** ↑	**ERR** ↓	**F1** ↑	**ERR** ↓	**F1** ↑
FSC	0.74	88.85%	Random (uniform)	67%	22%	75%	14%	80%	11%	83%	7%
			Random (majority)	10%	32%	13%	23%	16%	18%	21%	14%
			KNN	6%	78%	8%	66%	11%	60%	12%	62%
			CSI w/out CM pretrain	10%	32%	12%	27%	15%	25%	16%	26%
			CSI	**4%**	**88%**	**6%**	**77%**	**8%**	**75%**	**8%**	**77%**
LS	0.73	74.54%	Random (uniform)	67%	32%	75%	23%	79%	17%	83%	14%
			Random (majority)	56%	20%	60%	14%	62%	13%	67%	10%
			KNN	18%	68%	31%	50%	32%	50%	43%	37%
			CSI w/out CM pretrain	32%	53%	39%	47%	41%	40%	42%	30%
			CSI	**16%**	**83%**	**24%**	**58%**	**29%**	**54%**	**31%**	**50%**

5. RESULTS AND DISCUSSION

Table 1 shows the results of the CM error identification (Step 1, left) and the challenging subgroup identification (Step 3, right). We characterize the utterances via multiple features. Specifically, we use logits, speech metadata, and average hidden state values, and also the output probabilities for FSC and n-best sequence lengths for the LIBRISPEECH. We analyze the impact of each feature on performance. For FSC, the predominant factor for optimal performance is the output probabilities, whereas, for LIBRISPEECH, the hidden states.

We compare our methodology with three baseline approaches. In the first, class assignments to utterances are done randomly, while in the second, samples are assigned to the majority class. The third baseline is based on the K-nearest neighbors (KNN) classification. Specifically, we use as KNN representations the same input space as our CM-based approach. At inference time, samples are assigned to the closest K classes of training data points. The results underscore the efficacy of our proposed method, significantly outperforming the performance obtained by these strategies.

The CM trained to identify errors achieves an AUC of 0.74 for FSC and 0.73 for LIBRISPEECH. Although the AUC score may not reach exceptionally high levels, the pre-training phase of the CM proves its effectiveness in identifying particularly challenging subgroups within the data. We obtain high enhancements in the Error Rate (ERR) and F1 macro score of the CSI model by fine-tuning the pre-trained CM rather than training it anew (CSI w/out CM pretrain). Focusing on the FSC dataset and K=2, the two most negatively divergent subgroups consist of speech- and task-related metadata, i.e., {*num words=low, location=none, object=heat*} and {*trimmed duration=low, action=decrease*}. For this configuration, the F1 score increases from 32% to 88%, with a decrease in the Error Rate (ERR) from 10% to 4%. Similarly, we observe improvements for the LIBRISPEECH dataset, where the Error Rate drops from 32% to 16%, and the F1 macro score increases from 53% to 83%.

This trend holds across the investigated values of K, ranging from 2 to 5. In each case, the CSI fine-tuned from the confidence model consistently outperforms the model trained entirely from scratch. This suggests that the CM has indeed acquired valuable knowledge that proves beneficial for the task of identifying challenging subgroups. Our findings highlight the significance of pre-training the CM model, as it effectively contributes to identifying challenging subgroups.

6. CONCLUSIONS

We propose a novel Challenging Subgroup Identification model to predict the challenging subgroup to which an utterance belongs. The approach relies on fine-tuning a confidence model that captures error sources. The results show the effectiveness of our approach in identifying challenging subgroups without the need for demographics at deployment time.

We plan to further investigate our approach potential on more datasets (e.g., [3, 22]) and against stronger baselines (e.g., [3]), and consider a multi-label setting.

7. ACKNOWLEDGMENTS

This work is partially supported by 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.

8. REFERENCES

[1] Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke, "Toward fairness in speech recognition: Discovery and mitigation of performance disparities," in *Proc. Interspeech*. 2022, ISCA.

[2] Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Luca Cagliero, Luca de Alfaro, Elena Baralis, and Daniele 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.

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

[4] Yuanyuan Zhang, Yixuan Zhang, Bence Mark Halpern, Tanvina Patel, and Odette Scharenborg, "Mitigating bias against non-native accents," in *Interspeech 2022*, 2022.

[5] Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Elisa Reale, Luca Cagliero, Sandro Cumani, Luca de Alfaro, Elena Baralis, and Daniele Amberti, "Towards comprehensive subgroup performance analysis in speech models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[6] Irina-Elena Veliche and Pascale Fung, "Improving fairness and robustness in end-to-end speech recognition through unsupervised clustering," in *ICASSP 2023*, 2023.

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

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

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

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

[11] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio, "Speech model pre-training for end-to-end spoken language understanding," in *Interspeech 2019*, 2019, pp. 814–818.

[12] Eliana Pastor, Andrew Gavgavian, Elena Baralis, and Luca de Alfaro, "How divergent is your data?," *Proceedings of the VLDB Endowment*, vol. 14, no. 12, pp. 2835–2838, 2021.

[13] Eliana Pastor, Luca de Alfaro, and Elena Baralis, "Looking for trouble: Analyzing classifier behavior via pattern divergence," in *Proceedings of the 2021 International Conference on Management of Data*, New York, NY, USA, 2021, pp. 1400–1412.

[14] Woojay Jeon, Maxwell Jordan, and Mahesh Krishnamoorthy, "On modeling asr word confidence," in *ICASSP 2020*, 2020.

[15] Amber Afshan, Kshitiz Kumar, and Jian Wu, "Sequence-level confidence classifier for asr utterance accuracy and application to acoustic models," *Interspeech*, 2021.

[16] David Qiu, Qiujia Li, Yanzhang He, Yu Zhang, Bo Li, Liangliang Cao, Rohit Prabhavalkar, Deepti Bhatia, Wei Li, Ke Hu, et al., "Learning word-level confidence for subword end-to-end asr," in *ICASSP 2021*. IEEE, 2021, pp. 6393–6397.

[17] Qiujia Li, David Qiu, Yu Zhang, Bo Li, Yanzhang He, Philip C Woodland, Liangliang Cao, and Trevor Strohman, "Confidence estimation for attention-based sequence-to-sequence models for speech recognition," in *ICASSP 2021*.

[18] Qiujia Li, Yu Zhang, David Qiu, Yanzhang He, Liangliang Cao, and Philip C Woodland, "Improving confidence estimation on out-of-domain data for end-to-end speech recognition," in *ICASSP 2022,*.

[19] Jan Niehues and Ngoc-Quan Pham, "Modeling confidence in sequence-to-sequence models," *arXiv preprint arXiv:1910.01859*, 2019.

[20] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, 2020.

[21] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," 2021.

[22] Alkis Koudounas, Moreno La Quatra, Lorenzo Vaiani, Luca Colomba, Giuseppe Attanasio, Eliana Pastor, Luca Cagliero, and Elena Baralis, "ITALIC: An Italian Intent Classification Dataset," in *Proc. INTERSPEECH 2023*, 2023, pp. 2153–2157.