Prioritizing Data Acquisition For End-to-End Speech Model Improvement

PRIORITIZING DATA ACQUISITION FOR END-TO-END SPEECH MODEL IMPROVEMENT

Alkis Koudounas[†], Eliana Pastor[†], Giuseppe Attanasio[⋆], Luca de Alfaro[‡], and Elena Baralis[†]

[†]Politecnico di Torino, Turin, Italy, [⋆]Bocconi University, Milan, Italy,
[‡]University of California, Santa Cruz, CA, USA

ABSTRACT

As speech processing moves toward more data-hungry models, data selection and acquisition become crucial to building better systems. Recent efforts have championed quantity over quality, following the mantra "The more data, the better." However, not every data brings the same benefit. This paper proposes a data acquisition solution that yields better models with less data – and lower cost. Given a model, a task, and an objective to maximize, we propose a process with three steps. First, we assess the model's baseline performance on the task. Second, we use efficient mining techniques to identify subgroups that maximize the target objective if acquired first as new samples. Being the subgroups interpretable, we can determine which samples to acquire. Third, we run incremental training sampling from those subgroups. Experiments with two state-of-the-art speech models for Intent Classification across two datasets in English and Italian show that our method is significantly better than random or complete acquisition and clustering-based techniques.

***Index Terms*—** spoken language understanding, data acquisition, data markets, divergence

1. INTRODUCTION

Data has become increasingly important in building high-performing speech models. Deep learning-based speech models are data-hungry, relying on large amounts of data for effective training and optimization. Consequently, data selection and acquisition become critical to building better systems.

The default strategy is to prioritize the data volume, with the idea that the more data is used for training, the higher the performance of the resulting model. This strategy works up to a point, but has pitfalls with respect to model fairness, and, interestingly, even with respect to performance. For example, one may acquire data for already well-represented and modeled sub-populations, bringing little if any improvement to those sub-populations, and worsening the performance on sub-populations that tend to be under-represented in training data. More data does not entail better data. More data also means more cost in data collection, labeling, training, and verification. We show that for model improvement, a *targeted* data acquisition strategy aimed at compensating a model's weaknesses can work better than an *indiscriminate* strategy.

Specifically, we consider the problem of improving the quality of a trained model via data acquisition. We propose a *divergence-aware* acquisition approach, in which we leverage the techniques of [1] to automatically identify data subgroups on which model performance is inferior to the average. As the subgroups are interpretable, we propose to use them to guide data acquisition.

Previous data augmentation strategies generally rely on either a fixed set of user-defined subgroups, or on automatically created, but non-interpretable, subgroups. The former approach may miss unexpected problematic subgroups; the latter can be used to guide data augmentation but not data acquisition. In contrast, our approach can autonomously discover that a model struggles with, for example, utterances of women over 60 and guide data acquisition accordingly.

We assess the proposed divergence-guided acquisition on two models across two Intent Classification datasets in English and Italian, and we show that our proposed acquisition approach leads both to better performance, and better performance across subgroups (that is, model fairness) compared to an indiscriminate acquisition strategy. Further, we show that the advantage of the divergence-guided approach persists even when we acquire data that is a *subset* of that acquired by the indiscriminate strategy.

2. RELATED WORK

Prior works addressing data acquisition in speech generally focus on the diversity and robustness of the data, tackling the challenges from linguistic variations, recording conditions, environment, and demographics [2]. The considered groups are user-defined and known a priori. In contrast, our approach can discover problematic subgroups automatically, by exploring how model performance varies across combinations of interpretable attributes.

Given a set of groups of interest, recent work addressed the question of how many data samples from each group we should acquire to improve model performance by leveraging learning curves [3, 4]. The work in [5] automatically groups data by clustering speaker embeddings and identifies the clusters that exhibit inferior performance for a given model. Data

augmentation considers data samples close to the problematic clusters. However, these clusters, unlike our subgroups, are not interpretable. Subgroup interpretability is key for data acquisition. Interpretable subgroups allow the selection and acquisition of the required data from data vendors, or the planning of appropriate data acquisition strategies. Non-interpretable subgroups only allow filtering already-available data, e.g., by feeding it into an embedding model.

An approach that is close, and complementary, to ours is [6], which also proposes the automatic identification of subgroups defined by attribute combinations. Their work extracts subgroups that are under-represented in the training data and identifies the least amount of data to acquire to achieve adequate coverage. In contrast, we focus on identifying subgroups on which the model *under-performs*, so that more data can be acquired to specifically address this issue. Naturally, the two approaches can be used jointly.

3. METHODOLOGY

Our strategy to identify the subgroups with which the model struggles, and acquire data accordingly, entails two steps: (i) automatic under-performing subgroup detection, and (ii) divergence-aware data acquisition.

Automatic under-performing subgroup detection. Consider a set of labeled utterances and a speech model of interest. As a first step, we annotate the utterances with metadata. Metadata describe the utterances by means of interpretable features, such as demographic information of the speaker (e.g., gender, age, origin) and speaking and recording conditions (e.g., noise level, speaking rate, audio sample duration). It may be either already available in the dataset, or automatically derived [7] from utterances.

A subgroup is a subset of the utterances sharing the same metadata. We represent a subgroup as a conjunction of attribute-value pairs. For example, {gender=female, age>60} denotes the utterances of women with more than 60 years. Subgroups may overlap: for example, the previous subgroup overlaps with {gender=female}.

Once the utterances are annotated, we use DIVEXPLORER [1, 8] to extract all subgroups over metadata with adequate representation and estimate their model performance. By setting a frequency threshold, we ensure that the subgroups contain enough data instances to make the performance evaluation statistically significant. For each subgroup, we define its *divergence* Δ as the difference between the performance on the subgroup, and the performance on the entire dataset [1, 9]. Consider, for example, accuracy as a performance measure, and let S be a subgroup. A negative divergence in accuracy, denoted as $\Delta^{-}_{acc}(S)$, denotes a worse performance of the subgroup than the overall performance, and the lower the divergence, the more the model struggles in modeling it.

Divergence-aware Data Acquisition. Once assessed the performance of a given speech model, we aim to perform a data acquisition that may improve it, both in terms of overall performance, and across subpopulations. Here, we focus on accuracy as a performance measure, but other performance measures could be applied.

We let $\mathbb{S}^{-}$ be the set of *critical* subgroups, consisting of the subgroups that have negative divergence. Our subgroups have the key feature of being interpretable. Hence, we can target the acquisition of data samples with characteristics of the identified challenging subgroups.

We perform a pruning step to reduce redundancy among the critical subgroups, following the pruning approach defined in [1]. In this pruning, given a subgroup S_a and a subgroup S_b that includes S_a and another metadata condition, we only keep the more general S_a if the absolute difference in the divergence of the two subgroups is lower than a predefined threshold. The idea is that S_a already represents the divergence of S_b, as the additional features of S_b affect only slightly the divergence. For example, if the subgroup {young_woman} has divergence -0.39, and {young_woman, utterance>10s} has divergence -0.41, we keep only the former subgroup, as it accounts for most of the divergence of the latter. Pruning the critical subgroups results in a more concise representation, and facilitates data acquisition, as we can focus on the most relevant attributes.

We target for performance improvement the top-K summarized critical subgroups with the highest negative divergence, by acquiring data belonging to these subgroups. Specifically, we retrain the model with the addition of new data that belongs to one or more of the K subgroups (as subgroups can overlap, the same data instance can belong to more than one top-K subgroup). Thus, 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.

4. EXPERIMENTAL SETUP

We split our datasets into training, validation, held-out, and test sets. We use the same validation and test sets as in the original dataset. We split the training set, using 80% of it for training, and 20% held out. We use the validation set to identify the critical subgroups. We then acquire data samples from the held-out set and we retrain the model including the new samples. Finally, we evaluate the model performance, overall and on subgroups, using the test set. The code used in the paper can be found in [10].

Datasets. We evaluate our targeted acquisition approach on two datasets: Fluent Speech Commands (FSC) [11] for the English language and ITALIC [12] for Italian. FSC contains 30,043 utterances from 97 speakers, each with three labeled slots (action, object, location), which combined define the intent. ITALIC has 16,521 audio samples from 70 Italian

Table 1. Mean and standard deviation with three different runs for FSC dataset, wav2vec 2.0 base model. We compare the results for the *original* fine-tuning procedure, the two baselines (*random* and *clustering*-based) and *our* divergence-aware strategy. Best results for each number of considered subgroups K are highlighted in bold. Best results overall are underlined.

K	*Approach*	*#samples*	*Accuracy*	*F1 Macro*	Δ^-_{max}	Δ^-_{avg-10}	Δ^-_{avg-20}	Δ^-_{avg-50}	$\lvert\Delta_{avg-all}\rvert$
-	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
2	random	406	94.26 ± 0.27	91.17 ± 0.86	-54.26 ± 1.14	-53.93 ± 1.17	-53.24 ± 1.12	-52.37 ± 0.55	0.86 ± 0.06
	random	226	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
	clustering	406	92.94 ± 0.07	90.82 ± 1.19	-51.81 ± 0.86	-51.22 ± 0.92	-49.99 ± 0.10	-48.52 ± 0.11	1.24 ± 0.09
	clustering	226	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
	ours	226	**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**
3	random	874	92.21 ± 0.49	90.30 ± 0.55	-64.72 ± 3.07	-62.10 ± 2.49	-56.56 ± 2.32	-51.57 ± 1.88	0.38 ± 0.06
	random	382	94.13 ± 0.58	91.51 ± 0.82	-52.99 ± 3.40	-51.92 ± 3.02	-49.39 ± 2.21	-45.98 ± 1.78	0.33 ± 0.04
	clustering	874	**94.47 ± 0.44**	92.23 ± 0.45	-47.33 ± 0.33	-45.83 ± 0.16	-42.73 ± 0.21	-39.72 ± 0.49	0.32 ± 0.03
	clustering	382	90.03 ± 0.97	85.30 ± 0.94	-46.40 ± 0.36	-45.02 ± 0.33	-41.59 ± 0.28	-37.79 ± 0.16	0.81 ± 0.02
	ours	382	93.62 ± 0.29	**92.96 ± 0.46**	**-42.23 ± 0.12**	**-42.21 ± 0.11**	**-41.48 ± 0.11**	**-33.61 ± 0.07**	**0.22 ± 0.02**
4	random	1046	91.31 ± 0.98	89.48 ± 0.52	-61.85 ± 1.58	-60.72 ± 1.28	-58.08 ± 0.86	-54.83 ± 1.00	1.19 ± 0.03
	random	422	92.64 ± 0.27	91.29 ± 0.21	-55.83 ± 2.11	-55.71 ± 2.04	-51.41 ± 1.74	-45.41 ± 1.74	0.39 ± 0.02
	clustering	1046	93.28 ± 0.19	91.42 ± 0.18	-52.28 ± 0.63	-51.08 ± 0.58	-48.65 ± 0.40	-45.35 ± 0.44	0.85 ± 0.09
	clustering	422	87.72 ± 0.71	83.42 ± 0.48	-47.59 ± 0.25	-46.98 ± 0.21	-45.69 ± 0.12	-43.98 ± 0.09	0.72 ± 0.03
	ours	422	**95.16 ± 0.11**	**92.47 ± 0.22**	**-45.68 ± 0.24**	**-44.56 ± 0.25**	**-41.53 ± 0.24**	**-37.02 ± 0.20**	**0.15 ± 0.01**
5	random	1276	92.01 ± 0.49	91.00 ± 0.65	-67.77 ± 1.96	-66.94 ± 1.55	-65.31 ± 1.23	-62.65 ± 1.19	0.48 ± 0.03
	random	509	91.48 ± 0.55	90.27 ± 0.49	-54.82 ± 3.41	-54.75 ± 3.29	-54.69 ± 3.11	-51.12 ± 2.25	0.96 ± 0.08
	clustering	1276	92.75 ± 0.21	90.66 ± 0.22	-61.04 ± 0.19	-60.84 ± 0.24	-57.84 ± 0.18	-49.72 ± 0.11	1.33 ± 0.01
	clustering	509	91.44 ± 1.41	87.92 ± 1.38	-51.92 ± 0.19	-51.90 ± 0.24	-49.79 ± 0.18	-43.39 ± 0.11	0.45 ± 0.03
	ours	509	**94.12 ± 0.13**	**92.57 ± 0.16**	**-49.33 ± 0.15**	**-49.29 ± 0.12**	**-48.11 ± 0.21**	**-39.01 ± 0.11**	**0.11 ± 0.02**
-	all data	4606	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

speakers. The action and scenario slots determine the intent. We use the "Speaker" configuration, mirroring FSC's setup, with distinct speakers in the train, validation, and test sets.

Metadata. We enrich the datasets with various metadata following the approach described in [7].

Models. We perform fine-tuning on two end-to-end speech models: the wav2vec 2.0 [13] base model with approximately 90 million parameters on the FSC dataset and the multilingual XLSR model [14] with around 300 million parameters on ITALIC. We use the initial pre-trained model checkpoints available in the Hugging Face hub repository [15].

Metrics. We evaluate model performance using the accuracy and F1 Macro scores. We also evaluate performance at the subgroup level. We focus on the subgroup that shows the most substantial decrease in performance compared to the overall average, i.e., the highest negative divergence (Δ^-_{max}). We also compute the average divergence across the top 10, 20, and 50 subgroups with the highest negative divergence (Δ^-_{avg-n}), as well as the average absolute divergence across all identified subgroups ($|\Delta_{avg-all}|$).

Baselines. We evaluate the effectiveness of our method against two baseline approaches. The first baseline draws inspiration from [5] and employs an unsupervised clustering approach. For the baseline, we extract acoustic embeddings from the audio samples and group them into clusters. We select the clusters with the poorest performance and acquire the data points closest to those clusters from the hold-out set. Note that these clusters lack interpretability since the chosen data points are not selected based on annotated metadata, e.g., "utterances of young men speaking slowly", but rather on the distance measured in the embedding space. Thus, this technique can be used to select data samples from an available dataset, but it cannot be used to guide data collection or data acquisition. Given the specific characteristics of our datasets, we empirically determined that 20 clusters adequately capture the speech characteristics for FSC, while we used 10 clusters for ITALIC. Additional discussions on the cluster number are available in our project repository [10]. The second baseline approach involves adding instances selected at random from the held-out data set to the training data.

Experimental setting. For our approach, we acquire all the samples in the held-out set sharing the same metadata of the K critical subgroups. For the clustering-based baseline, we include in the training all the samples of the held-out set that have as closest-cluster one of the K clusters selected for improvement. This yields more data than with our targeted acquisition strategy (due to the narrowness of the latter). We compare targeted acquisition vs. clustering-based baseline both by allowing the clustering-based baseline to benefit from more data, and by assuming that targeted acquisition and clustering baseline can acquire the same amount of data. For the random baseline, we test two configurations, in which we randomly acquire a number of samples equal to the one acquired (i) with our technique and (ii) with the clustering one. We also report the results when we acquire all held-out data.

Table 2. Mean and standard deviation with three different runs for the ITALIC dataset and multi-lingual XLS-R-300.

K	*Approach*	*#samples*	*Accuracy*	*F1 Macro*	Δ^-_{max}	Δ^-_{avg-10}	Δ^-_{avg-20}	Δ^-_{avg-50}	$\lvert\Delta_{avg-all}\rvert$
-	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
2	random	383	75.34 ± 0.32	69.75 ± 0.59	-40.12 ± 1.47	-40.01 ± 1.46	-39.21 ± 1.33	-35.81 ± 0.84	0.37 ± 0.03
	random	154	75.32 ± 0.63	70.72 ± 0.58	-47.00 ± 0.81	-46.86 ± 0.80	-46.22 ± 0.77	-41.68 ± 0.70	0.38 ± 0.02
	clustering	383	74.35 ± 0.12	69.51 ± 0.30	-41.64 ± 0.60	-41.52 ± 0.60	-40.84 ± 0.52	-36.90 ± 0.38	**0.32 ± 0.02**
	clustering	154	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
	ours	154	**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
3	random	548	76.38 ± 0.12	71.09 ± 0.43	-40.51 ± 1.07	-40.41 ± 1.06	-39.52 ± 1.03	-35.12 ± 0.80	0.23 ± 0.04
	random	252	75.81 ± 0.13	71.46 ± 0.25	-51.32 ± 1.30	-51.14 ± 1.29	-50.27 ± 1.16	-45.37 ± 0.89	0.25 ± 0.01
	clustering	548	75.71 ± 0.12	71.31 ± 0.18	-39.74 ± 2.24	-39.65 ± 2.22	-38.81 ± 2.09	-35.02 ± 1.70	0.29 ± 0.01
	clustering	252	75.87 ± 0.20	70.70 ± 0.31	-42.93 ± 0.52	-42.82 ± 0.51	-41.89 ± 0.51	-37.25 ± 0.48	0.25 ± 0.02
	ours	252	**76.50 ± 0.30**	**71.69 ± 0.59**	**-36.73 ± 0.33**	**-36.57 ± 0.32**	**-36.18 ± 0.30**	**-32.20 ± 0.57**	**0.17 ± 0.03**
4	random	945	75.90 ± 0.30	70.83 ± 0.39	-42.34 ± 1.23	-42.23 ± 1.22	-41.65 ± 1.15	-37.83 ± 0.76	0.19 ± 0.02
	random	540	75.67 ± 0.20	71.71 ± 0.02	-41.07 ± 0.69	-40.96 ± 0.68	-40.36 ± 0.72	-36.41 ± 0.77	0.34 ± 0.05
	clustering	945	76.02 ± 0.40	71.53 ± 0.63	-42.52 ± 3.26	-42.43 ± 3.24	-41.76 ± 3.16	-37.97 ± 2.71	0.26 ± 0.06
	clustering	540	75.76 ± 0.21	71.22 ± 0.17	-41.50 ± 0.80	-41.40 ± 0.80	-40.79 ± 0.80	-37.87 ± 0.71	0.22 ± 0.03
	ours	540	**76.29 ± 0.13**	**72.48 ± 0.48**	**-37.30 ± 1.05**	**-37.22 ± 1.04**	**-36.79 ± 1.03**	**-33.42 ± 0.75**	**0.16 ± 0.04**
5	random	1035	**77.19 ± 0.34**	71.51 ± 0.39	-46.37 ± 0.88	-46.27 ± 0.89	-45.73 ± 0.98	-41.38 ± 1.11	0.21 ± 0.05
	random	604	75.13 ± 0.05	71.26 ± 0.17	-41.91 ± 1.95	-41.79 ± 1.94	-41.09 ± 1.83	-37.53 ± 1.26	0.29 ± 0.04
	clustering	1035	77.05 ± 0.22	**71.93 ± 0.04**	-42.93 ± 1.04	-42.86 ± 1.05	-42.39 ± 1.11	-38.34 ± 1.14	0.18 ± 0.03
	clustering	604	75.88 ± 0.29	70.60 ± 0.59	-40.41 ± 1.17	-40.32 ± 1.16	-39.47 ± 1.10	-36.12 ± 0.81	0.19 ± 0.03
	ours	604	77.14 ± 0.04	71.32 ± 0.41	**-37.52 ± 0.78**	**-37.44 ± 0.77**	**-36.83 ± 0.76**	**-33.75 ± 0.40**	**0.09 ± 0.02**
-	all data	2625	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

5. RESULTS AND DISCUSSION

Tables 1 and 2 compare the results of our approach and the baseline methods. Our targeted acquisition approach consistently demonstrates superior performance in terms of accuracy, F1 score, maximum divergence, and average divergence.

The best F1 score and accuracy performance for both datasets is observed when we selectively consider only the top-2 problematic subgroups (last row of the second block in both tables). Furthermore, the model also exhibits the lowest highest divergence (Δ^-_{max}) and the lowest average divergence for the top-10 (Δ^-_{avg-10}), top-20 (Δ^-_{avg-20}), and top-50 (Δ^-_{avg-50}) subgroups with the highest negative divergence. These findings suggest that an appropriate selection of a smaller set of samples can lead to significant performance improvements, both at the overall and subgroup levels.

As we increase the number of problematic subgroups for which we acquire data (i.e., K=3, 4, and 5), we notice a slight decrease in performance compared to K=2; the performance is nevertheless significantly better than the one of the original model (first row of the tables) and the one obtained when adding all available data (last row of the tables).

We observe a different trend in the overall average absolute divergence ($|\Delta_{avg-all}|$). The lower is $|\Delta_{avg-all}|$, the more the model shows less performance disparities across the subgroups. The lowest $|\Delta_{avg-all}|$ is consistently found for both datasets and models when K=5. This is intuitive, as adding more samples allows the model to address more problematic subgroups simultaneously. On the other hand, being the improvement more distributed across subgroups, we have the lowest improvement for the highest negative divergence.

6. CONCLUSIONS

We investigated a novel data acquisition approach to improve the performance of end-to-end speech models. Our results show that less data, acquired with the guide of subgroup divergence, can lead to higher performance than more data, indiscriminately acquired. Our approach outperforms the baseline methods both overall and at the subgroup level.

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), 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, and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 949944, INTEGRATOR). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

8. REFERENCES

[1] 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*. 2021, SIGMOD '21, p. 1400–1412, ACM.

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

[3] Irene Chen, Fredrik D Johansson, and David Sontag, "Why is my classifier discriminatory?," *Advances in neural information processing systems*, vol. 31, 2018.

[4] Ki Hyun Tae and Steven Euijong 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.

[5] 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*, 2022, pp. 1268–1272.

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

[7] 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.

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

[9] Eliana Pastor, Elena Baralis, and Luca de Alfaro, "A hierarchical approach to anomalous subgroup discovery," in *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 2023, pp. 2647–2659.

[10] Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Luca de Alfaro, and Elena Baralis, "Official code repository," [Online]: `https://github.com/koudounasalkis/Data-Acquisition-for-Speech-Model-Improvement`, January 2024.

[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, 20th Annual Conference of the International Speech Communication Association*, 2019, pp. 814–818.

[12] 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.

[13] 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, vol. 33, pp. 12449–12460.

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

[15] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, Oct. 2020, pp. 38–45, Association for Computational Linguistics.