Exploring Subgroup Performance In End-to-End Speech Models

Original
Exploring Subgroup Performance In End-to-End Speech Models / Koudounas, Alkis; Pastor, Eliana; Attanasio, Giuseppe; Mazzia, Vittorio; Gioletto, Manuel; Gueudre, Thomas; Caglieri, Luca; de Alfaro, Luca; Baralis, Elena Maria; Amberti, Daniele. - (2023), pp. 1-5. (Intervento presentato al convegno 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) tenutosi a Rhodes Island (Greece) nel 4-10 June 2023) [10.1109/ICASSP49357.2023.10095284].

Availability:
This version is available at: 11583/2976783 since: 2023-03-13T10:48:09Z

Publisher:
IEEE

Published
DOI:10.1109/ICASSP49357.2023.10095284

Terms of use:
This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)
EXPLORING SUBGROUP PERFORMANCE IN END-TO-END SPEECH MODELS

Alkis Koudounas♢∗, Eliana Pastor♢∗, Giuseppe Attanasio♡, Vittorio Mazzia♣, Manuel Giollo♣
Thomas Gueudre♣, Luca Cagliero♢, Luca de Alfaro♢, Elena Baralis♢, Daniele Amberti♣

♢Politecnico di Torino, Turin, Italy, ♣Alexa AI, Amazon, Turin, Italy
♡Bocconi University, Milan, Italy, ♢University of California, Santa Cruz, CA, USA

ABSTRACT

End-to-End Spoken Language Understanding models are generally evaluated according to their overall accuracy, or separately on (a priori defined) data subgroups of interest. We propose a technique for analyzing model performance at the subgroup level, which considers all subgroups that can be defined via a given set of metadata and are above a specified minimum size. The metadata can represent user characteristics, recording conditions, and speech targets. Our technique is based on advances in model bias analysis, enabling efficient exploration of resulting subgroups. A fine-grained analysis reveals how model performance varies across subgroups, identifying modeling issues or bias towards specific subgroups.

We compare the subgroup-level performance of models based on wav2vec 2.0 and HuBERT on the Fluent Speech Commands dataset. The experimental results illustrate how subgroup-level analysis reveals a finer and more complete picture of performance changes when models are replaced, automatically identifying the subgroups that most benefit or fail to benefit from the change.

1. INTRODUCTION

End-to-End Spoken Language Understanding (E2E SLU) models achieve state-of-the-art performance on Natural Language Understanding tasks without converting speech into the underlying text. Speech data often comes with additional information about the speaker (e.g., the age), recording conditions (e.g., the noise level), or task characteristics (e.g., the uttered intent), among other things. We define this information as speech metadata. Combinations of metadata values identify data subgroups. Typically, model performance is evaluated either on the whole testing set or on relevant data subgroups identified in advance.

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Sup</th>
<th>acc</th>
<th>Δ_acc</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>{age=22-40, gender=male, loc=none, speakRate=high, tot_silence=high}</td>
<td>0.03</td>
<td>74.79</td>
<td>-18.38</td>
<td>4.7</td>
</tr>
<tr>
<td>{action=increase, gender=male, speakRate=high}</td>
<td>0.03</td>
<td>74.81</td>
<td>-18.36</td>
<td>4.9</td>
</tr>
</tbody>
</table>

Table 1. wav2vec 2.0 large accuracy gap (Δ_acc) for two identified subgroups compared to overall test accuracy.

We introduce efficient techniques for comparing model performance on all data subgroups that are induced by the available metadata. Since the number of subgroups is exponential in the number of metadata attributes, the naive enumeration and evaluation of subgroups is unfeasible. Our approach leverages advances in model bias analysis [1]. The basic insight is that while the number of subgroups is exponential, the number of subgroups above a specified size (for instance, containing at least 0.1% of the dataset) is generally not. These subgroups are called the frequent subgroups: they are the subgroups with both practical and statistical significance. Our approach allows measuring and comparing model performance on all frequent subgroups. Among other things, this enables exploring the impact of sensitive attributes such as gender in isolation or in conjunction with other attributes. Table 1 reports an example of problematic subgroups where we find the model underperforming compared to overall statistics.

Our main contributions are as follows. First, we describe how to study the subgroup-level performance of speech E2E models, and we identify data subgroups on which a single model performs better or worse than average. Second, we extend the approach to the comparison of models, and we identify the subgroups on which performance most improves, or suffers, when a model is replaced with another. Lastly, we benchmark our proposed approach on the wav2vec 2.0 [2] and HuBERT [3] models, and the Fluent Speech Commands dataset [4]. Our approach can easily identify performance imbalances across subgroups defined by demographic features. We further show that an increase in model size and complexity does not necessarily yield a mitigation of model bias.
2. RELATED WORK

Prior works [5, 6, 7, 8, 9, 10, 11] have studied the presence of model bias and unfairness in data subgroups, mainly considering gender, accents, or age features. These related works are rooted in specific combinations of features (e.g., age, gender or skin tones [9], demographics, and geolocation [11]). In [11], the authors also propose to identify under-performing subgroups automatically by clustering speaker embeddings. However, subgroups are not interpretable. Differently, we slice over metadata, thus allowing their direct interpretation.

Recently, several works addressed the automatic identification of subgroups with anomalous behaviors on structured data [1, 12, 13, 14]. Our approach, which deals with speech data, is based on DIVEXPLORER [1] and proposes a method to allow model comparison on subgroups. While the heuristic-driven exploration approaches of [13, 14] do not support model comparison, DIVEXPLORER [1] is the only one that can support it because of its exhaustive exploration of frequent subgroups.

3. MEASURING SUBGROUP BEHAVIOR

We define data subgroups via itemsets, which are sets of attribute = value pairs. For a speech recognition model, the divergence of a data subgroup is the difference between the model performance on the subgroup and the model performance on the whole dataset [1]. Given two models, the subgroup gain is the difference in performance between the two models on the subgroup.

Datasets, metadata, and items. We annotate speech data with metadata consisting of interpretable attributes. They describe speaker-related features such as gender or age, and speaking and recording features, such as type of environment, presence and type of noise, and speaking rate. We also collect task-specific features, e.g., an intent description.

We denote by $D$ our dataset, by $A$ its set of metadata attributes, and by $Z$ its set of items: an item has the form $a = v$, for an attribute $a \in A$ and a value $v$. If gender and age are attributes, examples of items are gender = female and age $\in [20–40]$. The subgroup corresponding to an item is the portion of the dataset that satisfies it. For each attribute, we require that the item subgroups form a partition of the dataset. For instance, for the age attribute, the age ranges need to be non-overlapping, and their union must cover all possible ages.

An item enables us to slice, or select, a subset of the data with respect to one attribute. We can also slice the data with respect to multiple attributes by considering itemsets, which are collections of zero or more items, each item referring to a distinct attribute. An example of itemset is \{gender = male, age $\in [10, 20]$\}. For an itemset $I$, we let the support of $I$ be the fraction of the dataset that corresponds to $I$, that is, the ratio between the size of the subgroup satisfying $I$ and the size of the whole dataset. Thus, an itemset with support of 0.02 will appear in 2% of the dataset. The empty itemset corresponds to the entire dataset and has support 1.

Subgroup divergence and gain. Let $f$ be a performance measure for a downstream SLU task, so that for a model $M$ and a subgroup (i.e., itemset) $I$, $f(I, M)$ is the performance of the model on the subgroup. The performance can reflect correctness, top-$n$ correctness, or other standard measures of model performance. The divergence of itemset $I$ with respect to model $M$ is the difference between the model performance over $I$, and the one over the whole dataset [1]:

$$div_f(I, M) = f(I, M) - f(\emptyset, M).$$

We define the gain from model $M_1$ to model $M_2$ for itemset $I$ as the increase in performance on $I$ obtained by replacing model $M_1$ with model $M_2$:

$$gain_f(I, M_1, M_2) = f(I, M_2) - f(I, M_1).$$

We leverage DIVEXPLORER [1] to identify itemsets with large absolute-value divergence or gain. DIVEXPLORER integrates frequent pattern mining techniques to efficiently extract all itemsets above a given support threshold together their divergence. The support threshold binds the exploration and ensures that the returned itemsets contain sufficient data to be statistically and operationally significant.

We are also interested in characterizing the role of items in yielding itemsets with high divergence or gain. Let $g(I)$ be the metric of interest for itemsets ($g$ can be divergence or gain). Following [1], we define the contribution of $i \in I$ to $g(I)$ using the game-theoretical notion of Shapley value, attributing to a “team” $J \subseteq I$ of items the contribution $g(J)$. The Shapley value $s_g(i, I)$ of $i$ in $I$ captures the notion of how much $i$ contributed to the divergence or gain of $I$, and we have $\sum_{i \in I} s_g(i, I) = g(I)$. We also consider the global Shapley value $S_g(i)$ of an item $i$, which measures the average effect of adding item $i$ to all other compatible itemsets [1].

4. EXPERIMENTAL RESULTS

We evaluate the performance of our approach by showing its ability to reveal sources of error (§4.2), analyzing how model size impacts performance at the subgroup level (§4.3), comparing subgroup behavior across different models (§4.4), and studying the attribute role (§4.5).

4.1. Data and Models

Dataset. FLUENT SPEECH COMMANDS (FSC) [4] is one of the most used datasets for the Intent Classification (IC) task. We analyze the test set containing 3793 audio samples from 10 speakers. Each audio sample is associated with three slots: action, object, and location. The intent is the combination

\footnote{Code at https://github.com/dbdng/divergence-in-speech-systems.}
We investigate the sources of errors for wav2vec 2.0 large model. Table 1 reports the two itemsets with the highest negative divergence. These define the subgroups on which the model performs worse than the average. The groups consist of male speakers with a high speaking rate and either (i) age between 22 and 40, high duration of silences and “none” location (1st subgroup, accuracy lower by 18.38%), or (ii) increase action (2nd subgroup, accuracy lower by 18.36%).

**4.2. Individual Model Debugging**

Table 2. Performance gain comparing wav2vec 2.0 base to large (top) and wav2vec 2.0 base to HuBERT base (bottom) on itemsets where performance increases (↑), decreases (↓), or remains equal (=).

<table>
<thead>
<tr>
<th>Subgroups</th>
<th>Sup gainaccc_w2v2-b acc w2v2-l acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>↑ {action=increase, duration=low, loc=none, speakRatetrim=low, trim_dur=low}</td>
<td>0.03 22.69 75.63 98.32</td>
</tr>
<tr>
<td>= {action=increase, gender=male, n_words=low, speakRate=high}</td>
<td>0.03 0.0 75.41 75.41</td>
</tr>
<tr>
<td>↓ {action=activate, gender=male, speakRate=low}</td>
<td>0.03 -20.97 96.77 75.81</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subgroups</th>
<th>Sup gainaccc_w2v2-b acc hub-b acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>↑ {gender=male, loc=none, n_words=low, tot_silence=high, trim_dur=low}</td>
<td>0.03 31.20 64.00 95.20</td>
</tr>
<tr>
<td>= {action=decrease, n_words=high, speakRate_trim=medium}</td>
<td>0.04 0.0 95.49 95.49</td>
</tr>
<tr>
<td>↓ {action=decrease, age=22-40, loc=washroom}</td>
<td>0.03 -1.68 100.00 98.32</td>
</tr>
</tbody>
</table>

Table 3. Impact of gender on accuracy for wav2vec 2.0 base and large, increase (↑), decrease (↓) or equal (=) performance.

<table>
<thead>
<tr>
<th>Subgroups</th>
<th>gender=female</th>
<th>gender=male</th>
</tr>
</thead>
<tbody>
<tr>
<td>{action=increase, loc=none, n_words=low, trim_dur=low}</td>
<td>68.29 88.62 (↑) 73.48 81.82 (↑)</td>
<td>78.41 90.91 (↑) 75.41 75.41 (=)</td>
</tr>
<tr>
<td>{action=increase, n_words=low, speakRate=high}</td>
<td>96.17 95.22 (↑) 94.54 78.14 (↓)</td>
<td></td>
</tr>
<tr>
<td>{action=activate, n_words=medium}</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**4.3. Impact of Model Scale**

We limit gender to a binary variable under the current data regime.

Generally, larger models can be more accurate. Our analysis shows that even when the overall performance increases with a larger model, there can be subgroups where it decreases.

We compare the performance of wav2vec 2.0 base (w2v2-b) and large (w2v2-l) models. Overall, the accuracy rises from 91.72% for the smaller model to 93.17% for the larger one. To analyze subgroup performance, we compute the gain as the difference in performance between the smaller and larger models on the subgroup. Performance increases in 63.75% of the explored subgroups and decreases in 31.89% of them. Table 2 shows subgroups on which performance increases (↑), does not change significantly (=), and decreases (↓). The largest drop in performance (-20.97%) corresponds to male speakers, action activate, and low speaking rate.

Unequal improvement across genders. Our approach enables the comparison of model performance on any combination of metadata attributes. One such attribute is gender. Considering gender alone, the accuracy for women rises from 92.51% to 95.39%, while for men is almost stable, with a slight decrease (-0.11%). Consequently, the performance gap among genders increases when going to the larger model.
Further, our analysis reveals that this difference is exacerbated if we consider gender intersected with other factors. Table 3 compares the performance of the base vs. large model separately for the two main genders. The table reports the results for three sample itemsets. The small model performs worse for females in the first subgroup, while the larger shows better performance. Conversely, the larger model shows a significant drop in performance for males in the last subgroup.

4.4. Cross-Model Divergence

In the previous results, we have compared models that are of different sizes but share the same structure. We can use our techniques for comparing models with different structure as well. As an example, we can study the performance gain obtained by changing the wav2vec 2.0 with the HuBERT base model. Overall, Hub-b has higher performance than w2v2-b (98.42% compared to 91.72%). The second block of Table 2 reports results for three specific subgroups. The former is associated with the largest performance gain, for the second the performance is unchanged, and the latter is associated with the largest decrease. When changing from wav2vec 2.0 to HuBERT base, 97.03% of the subgroups have positive gain. When changing from base to large wav2vec 2.0, only 63.75% of subgroups had positive gain. Changing architecture to HuBERT has a far more uniformly positive effect on performance than increasing the wav2vec 2.0 model size. Our approach enables the study of how uniform the gains are when a model is replaced with a different one.

4.5. Attribute Role in Subgroup Behavior

Given a subgroup, it is interesting to understand the role that each metadata attribute plays in the divergence or gain of the subgroup. This role is captured by the notion of Shapley value (see Section 3 and [1]). Consider the first row of Table 3. It defines two subgroups, one for gender male, one for gender female, and both with \{action=\textit{increase}, loc=none, n_words=low, trim_dur=low\}. Figure 1 shows the influence of each item on the gain obtained by going from small to large models. We note that all four items \textit{action=increase}, \textit{loc=none}, \textit{n_words}=low, \textit{trim_dur}=low contribute positively to the performance gain. In contrast, item \textit{gender=female} has a positive influence for the female subgroup, while \textit{gender=\textbf{male}} has negative influence for the male subgroup. That is, the lesser performance gain for the male subgroup is chiefly caused by the \textit{gender=male} item.

\textbf{Global item role in performance gain.} We summarize the impact of each item on the gain in performance when the size is increased using the global Shapley value $S_g$. Intuitively, a positive value for $S_g(i)$ indicates that on average, adding $i$ to an itemset $J$ with $i \notin J$ leads to a performance increase; similarly for negative values. The top 15 items with the largest impact when changing from wav2vec 2.0 base to HuBERT base (shaded blue) are shown in Figure 2 (in blue). Utterances with low duration and a low number of words are associated with a performance improvement, while medium duration and the action \textit{activate} are associated with a decrease. The presence of gender among the top 3 confirms its role in model performance: other things being equal, the gain is greater for females and smaller for males. Figure 2 reports (in blue) the global Shapley value of the top 15 items when changing from wav2vec 2.0 base to HuBERT. For this model change, gender has only a small and negligible contribution to the gain.

5. CONCLUSIONS AND FUTURE WORK

We proposed a novel methodology to characterize the behavior of E2E speech representation models in data subgroups. We investigated subgroup behavior when the model’s size is increased, and a different model is adopted. We also revealed disparate improvement for sensitive attributes such as gender.

As the method is model- and task-agnostic, we envision extending its adoption to multiple models and tasks.

6. ACKNOWLEDGMENTS

This work is partially supported by the FAIR - Future Artificial Intelligence Research funded by the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013), the grant “National Centre for HPC, Big Data and Quantum Computing”, CN000013 (approved under the M42C Call for Proposals - Investment 1.4 - Notice “Centri Nazionali” - D.D. No. 3138, 16.12.2021, admitted for funding by MUR Decreto No. 1031,17.06.2022), Fondazione Cariplo (grant No. 2020-4288),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


[6] Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse 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.


