Towards Comprehensive Subgroup Performance Analysis in Speech Models

Alkis Koudounas *Graduate Student Member, IEEE,* Eliana Pastor *Member, IEEE,* Giuseppe Attanasio, Vittorio Mazzia, Manuel Giollo *Member, IEEE,* Thomas Gueudre, Elisa Reale, Luca Cagliero *Member, IEEE,* Sandro Cumani Luca de Alfaro, Elena Baralis *Member, IEEE,* and Daniele Amberti

***Abstract*—The evaluation of spoken language understanding (SLU) systems is often restricted to assessing their global performance or examining predefined subgroups of interest. However, a more detailed analysis at the subgroup level has the potential to uncover valuable insights into how speech system performance differs across various subgroups.**

In this work, we identify biased data subgroups and describe them at the level of user demographics, recording conditions, and speech targets. We propose a new task-, model- and dataset-agnostic approach to detect significant intra- and cross-model performance gaps. We detect problematic data subgroups in SLU models by leveraging the notion of subgroup divergence. We also compare the outcome of different SLU models on the same dataset and task at the subgroup level. We identify significant gaps in subgroup performance between models different in size, architecture, or pre-training objectives, including multi-lingual and mono-lingual models, yet comparable to each other in overall performance. The results, obtained on two SLU models, four datasets, and three different tasks—intent classification, automatic speech recognition, and emotion recognition—confirm the effectiveness of the proposed approach in providing a nuanced SLU model assessment.

***Index Terms*—Speech representation, E2E-SLU models, Subgroup identification, Model bias analysis, Divergence**

I. Introduction

SPEECH and language technologies have advanced significantly over the years, enabling the development of intelligent systems that can recognize, transcribe, and understand speech. These systems find applications in virtual assistants [1], customer service [2], healthcare [3], speech emotion recognition [4], and more. However, system evaluations often focus on overall performance, neglecting performance disparities across subgroups. Furthermore, the rise of large self-supervised pre-trained neural network models [5], characterized by larger size and complexity, significantly hampers interpretability and amplifies the challenges in accurately assessing capabilities and identifying potential performance inequalities. Thus, a comprehensive evaluation framework is necessary to capture nuances and ensure equitable performance assessment in speech and language technologies.

Recent studies revealed model bias and disparate treatment in data subgroups ([6], [7], [8], [9], [10], [11], [12], [13], [14]). A data subgroup is a subset of the data sharing some properties, such as similarity in the embedding space or common feature values (e.g., utterances of *female* speakers). Prior works generally focus on predefined subgroups defined by protected and known features of interest. However, identifying subgroups typically requires human expertise and often involves analyzing each attribute separately, which limits the exploration of unexpected and crucial subgroups. Our study introduces an automated approach for identifying critical subgroups to address these limitations. Unlike existing methods, that rely on clustering speaker embeddings [6], our approach allows for intersectional analysis, enabling the examination of combined effects across multiple attributes. 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). Other information, such as speaking rate and number of words, can be readily derived from the speech or transcripts. By combining metadata values, we can identify data subgroups. The subgroups generated through our method are easily understandable by humans, addressing the interpretability challenge often faced by existing automated methods.

Research goal. In this work, we study the presence of bias in spoken language understanding (SLU) model performance on data subgroups. We automatically identify those combinations of metadata values yielding maximal:

- *Intra-model performance gap*, i.e., a significant difference in performance between the overall dataset and the data subgroup. We quantify it by means of an established divergence metric [15] or
- *Cross-model performance gap*, i.e., a significant gap in the subgroup performance of models different in size, architecture, or pre-training objectives.

Evaluating intra-model subgroup divergence allows a more nuanced analysis of subgroup performance within a specific SLU model, whereas estimating cross-model gaps can guide end-users in choosing the best SLU model to use on a proprietary dataset.

A. Koudounas, E. Pastor, L. Cagliero, S. Cumani, and E. Baralis are with the Politecnico di Torino, Turin, Italy, e-mail: {name.surname}@polito.it.

G. Attanasio is with the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis, Milan, Italy.

V. Mazzia, M. Giollo, T. Gueudre, E. Reale, and D. Amberti are with AGI, Amazon, Turin, Italy.

L. de Alfaro is with the University of California, Santa Cruz, CA, USA.

This paper includes a supplementary that provides additional results on the whole set of datasets, tasks, and models for all the research questions addressed in our work. Contact `alkis.koudounas@polito.it` for further questions about this work.

Research questions. We aim to answer the following questions:

- **RQ1**. How can we automatically identify and describe the most problematic subgroups for a given combination of SLU model, dataset, and task?
- **RQ2**. What is the effect of the model size on subgroup performance? Does *the larger, the better* hold true?
- **RQ3**. Are the performance disparities on specific subgroups independent of the model architecture?
- **RQ4**. Are multilingual SLU models more sensitive to subgroup performance disparities than monolingual ones?

Running example. Let I=$\{$*trimmed speaking rate*=high, *gender*=female, *total duration*=low$\}$ be a conjunction of metadata values extracted from the LIBRISPEECH benchmark dataset [16]. I indicates a data subgroup consisting of short-lasting speeches made by female speakers at a relatively high speaking rate (words per second). Regarding the intra-model performance gap, an end-to-end user would like to analyze the WER performance of an established transformer-based SLU model, i.e., *wav2vec 2.0 base* [17] on the given data subgroup I for the ASR task. The WER of the global model (6.06%) is significantly better than that achieved on I (17.03%). The divergence of I, given by the performance difference (10.97%), indicates the presence of a potential bias. Let us now compare the WER of two different wav2vec 2.0 versions, i.e., *base* and *large*, on I. The cross-model WER gap between wav2vec 2.0 base (17.03%) and its large version (11.31%) is significant, showing a clear benefit in using a larger model.

Proposed approach. We present a novel methodology for automating the characterization and comparison of metadata-generated subgroups. The number of subgroups grows exponentially with the number of metadata attributes. Hence, it becomes infeasible to enumerate and evaluate them using naive approaches. Our proposed approach leverages recent advancements in model bias analysis to address this challenge [15], [18]. The critical insight lies in recognizing that, although the number of subgroups is exponential, the number of subgroups that exceed a certain coverage threshold (e.g., containing at least 0.1% of the dataset) is generally manageable. These subgroups, called "frequent subgroups", possess practical and statistical significance. On top of the generated patterns, we shortlist the subgroups with maximal intra- and cross-model gaps. They respectively provide end-users with explainable representations of problematic subgroups within a given SLU task and across different (but comparable) SLU models.

The main paper contributions, hereafter denoted by C1-5, are summarized below:

C1) *Problematic subgroup definition and error analysis.* We propose a new approach to explain SLU models at the level of data subgroups based on the concepts of intra- and cross-model performance gaps.

C2) *Impact of the size on model behavior.* Generally speaking, larger models will likely be more accurate and fair [19]. While the overall performance of a model likely increases as it is scaled up, our analysis shows that there may be subgroups where it unexpectedly decreases.

C3) *Impact of multi-lingual pre-training objective on model behavior.* Given the emerging trend of switching to multilingual models, we leverage our approach to explore advantages and disadvantages at the subgroup level when going from mono- to multi-lingual models.

C4) *Impact of the architecture on model behavior.* Our methodology can also be applied to compare models with different structures.

C5) *Datasets and models benchmarking.* We thoroughly analyzed four speech datasets, three tasks, and two models with two different sizes to discover and study models' behavior on specific subgroups. We conducted experiments on LIBRISPEECH [16] for Automatic Speech Recognition (ASR), FSC [20] and SLURP [21] for Intent Classification (IC), and IEMOCAP [22] for Emotion Recognition (ER), using two transformer-based models in base and large sizes, namely wav2vec 2.0 [17] and HuBERT [23]. We measured the accuracy and word error rate (WER) metrics and assessed their divergence across subgroups.

A preliminary version of the present work, focused on the intent classification task, has been presented in [24]. We analyzed the wav2vec 2.0 large model at the subgroup level and the impact of size (when changing from wav2vec 2.0 base to large) and architecture (when transitioning from wav2vec 2.0 to HuBERT base) for a single dataset, FSC [20]. This paper proposes a more comprehensive analysis across various speech datasets and tasks, including intent classification, emotion recognition, and automatic speech recognition. We analyze the performance at the subgroup level of two models, wav2vec 2.0 and HuBERT, in their two base and large sizes for four datasets. We further investigate how the size and architecture affect the overall and subgroup model performance, and we also study the impact of multilingual pre-training objectives.

To foster the reproducibility and dissemination of our research work, the source code and its documentation are available at `https://github.com/koudounasalkis/Subgroup-Analysis-in-Speech-Models`.

The remainder of this paper is organized as follows. Section II reviews the related works. Section III describes the methodology. Section IV reports the main experimental results. Finally, Section V draws conclusions and future directions.

II. RELATED WORK

The study of bias and fairness in speech models has gained significant attention recently. A growing body of work (e.g., [6], [7], [8], [9], [10], [11], [12], [13], [14]) has focused on identifying, measuring, and possibly mitigating the existence of model bias and unfairness in data subgroups, particularly on features such as gender, accents, or age.

Prior works focused on identifying bias in specific demographics metadata, such as the skin tone [7], the ethnicity [13], or in specific combinations of metadata, e.g., demographics and geolocation [6], [8], gender and ethnicity [10], gender, age, and accents [9] or gender, age, skin tones [11]. In [14], the authors study how to detect and mitigate ASR performance also for dysarthric speakers.

To the best of our knowledge, the first attempt to automatically identify arbitrary speech data subgroups was made in [6]. They identify underperforming speaker subgroups via speaker embeddings' clustering. Clusters in the latent space, however, are generated directly from the raw data and thus are not easily explainable. Unlike previous approaches, our work focuses on identifying problematic subgroups consisting of arbitrary metadata combinations. Differently from [6], we rely on explainable patterns consisting of conjunctions of metadata values and possibly covering non-disjoint speaker groups.

III. Methodology

Our approach analyzes model performance at the subgroup level. A subgroup is a subset of the data characterized by a set of metadata and their value, denoted in our paper as itemsets or patterns. The metadata can represent user characteristics (e.g., gender, age), speech targets (e.g., speaking rate, duration of silence), and dataset-related features (e.g., intents, labels). In our work, for example, the subgroup {*gender=female, age∈[20-40]*} indicates the subset of utterances of female speakers in the age 20-40.

We inspect subgroup behavior via two complementary notions, i.e., the intra-model divergence and cross-model gap:

- *Intra-model divergence* is the difference in model performance on a data subgroup and the whole dataset. We use this notion to inspect the behavior of an individual model to reveal which subgroups are associated with lower-than-average performance (but also higher or equal).
- *Cross-model performance gap* is the difference in performance between the two models on the same subgroup. It is used to compare different models at the level of subgroups.

The exploration of subgroups and the computation of their performance and divergence are efficiently performed via a lattice-based method.

The following subsections analyze our approach in detail. Specifically, Section III-A outlines the notion of slicing via interpretable metadata, Section III-B defines subgroup divergence and gap, and Section III-C describes the evaluation of the local and global contribution to divergence and gap via game theory concepts. Finally, Section III-D illustrates the DivExplorer algorithm that we leverage for an efficient subgroups exploration.

A. Slicing via interpretable metadata

We aim to characterize and understand speech model behavior in terms of interpretable data subgroups. We define interpretable metadata as data that humans can easily understand, such as the age or gender of the speaker or the utterance level of noise. We leverage such metadata to define subgroups. For example, a directly interpretable subgroup is the *young women in a noisy environment* utterances. In the following, we first illustrate the notion of metadata as interpretable attributes to enhance speech data. We then illustrate the process of slicing via metadata to define subgroups.

Metadata. We annotate speech data with metadata consisting of interpretable attributes. Metadata describes utterances in a human-understandable manner. Below, we report a (partial) characterization of what metadata can describe:

- *speaker demographics* such as gender or age;
- *speaking features*, e.g., the speaking rate and the duration of silences;
- *recording conditions*, such as type of environment and presence and type of noise;
- *task- or dataset-specific features*, e.g., an intent description for an intent classification task.

Metadata can be already available in the dataset under analysis, such as demographic information of the speakers or the target intent in a labeled dataset. We can also derive metadata from utterances or their transcriptions, such as the utterance duration, number of words, or speaking rate as the ratio of words per second.

We denote by D our dataset under analysis and by A its set of metadata attributes.

Slices, items, and itemsets. An *item* is an attribute equality with 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. We require that the item subgroups form a dataset partition for each attribute. For instance, the age ranges must be non-overlapping for the *age* attribute, and their union must cover all possible ages. For an attribute $a \in A$, we denote its number of possible values with m_a.

An item enables us to *slice*, or select, a subset of the data concerning one attribute. We can also slice the data concerning multiple attributes by considering *itemsets*, which are collections of zero or more items, each referring to a distinct attribute. An example of itemset is $\{gender = male, age \in [10, 20]\}$. We define data subgroups via *itemsets*, allowing for an interpretable definition of subgroups. We denote by $attr(I)$ the set of attributes included in an itemset I. For an itemset I, we let the *support* of I be the fraction of the dataset corresponding 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, denoted by $\emptyset$, corresponds to the entire dataset and has support 1. We say that an itemset is *frequent* with respect to a minimum support threshold u if its support is greater or equal to u.

For a subset of attributes $B \subseteq A$, we denote by $\mathcal{I}_B = \{I|\ attr(I) = B\}$ the itemsets over attributes B and by $\mathcal{I}_A$ the itemsets that contain all attributes of D. By $\mathcal{I}_B^{\star,u}$, we denote the set of frequent itemsets with attributes B for support threshold u. We will use $\mathcal{I}_B^{\star}$ when u is clear from the context.

B. Subgroup Divergence and Performance Gap

We are interested in identifying subgroups with different performances than the overall dataset. We use the notion of subgroup divergence as the difference in the performance of a subgroup compared to the whole dataset [15]. Evaluating differences at the subgroup level is also critical when comparing models. We introduce the notion of cross-model performance

gap as the difference in performance between the two models on the subgroup.

Subgroup divergence. Let f be a generic statistic for a downstream SLU task so that for a model M and a subgroup (i.e., itemset) I, $f(I, M)$ is the average of the statistic of the model on the subgroup. The statistic 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 [15]:

$$\Delta_f(I, M) = f(I, M) - f(\emptyset, M) \ . \tag{1}$$

The higher the divergence (in absolute terms), the more the performance in the subgroup diverges from the overall behavior. Consider, for example, the accuracy as the statistic f. A subgroup with negative (and high) divergence indicates that the model is underperforming. We use Welch's t-test, as outlined in [15], to determine the statistical significance of divergence. Our hypothesis tests whether the means of the subgroup I and the entire population D are equal for the statistic f.

Cross-model performance gap. We define the performance gap from model M_1 to model M_2 for itemset I as the change in performance on I obtained by replacing model M_1 with model M_2:

$$gap_f(I, M_1, M_2) = f(I, M_2) - f(I, M_1) \ . \tag{2}$$

The definition of subgroup divergence and performance gap can apply to generic SLU models for a generic task to assess the subgroup performance of a statistic f of a generic dataset annotated via metadata. This makes the methodology model-, task-, and metric- agnostic. To assess the statistical significance of performance gaps, we again use Welch's t-test. We test the hypothesis that the means of statistic f for models M_1 and M_2 are equal.

C. Local and global contribution to divergence or gap

Once we identify the itemsets with significant divergence or gap, it is interesting to characterize the role of their items in their divergence or gap. We use notions from game theory to provide a local and global understanding of the subgroup behavior.

Local contribution. Given an itemset I, the local contribution quantifies the local role of each item to its gap or divergence. Let $g(I)$ be the metric of interest for itemset I (g can be divergence or gap). Following [15], we define the contribution of $i \in I$ to $g(I)$ using the game-theoretical notion of *Shapley value*. The Shapley value assigns each team member their contribution to the team's total score. Paralleling its definition, we consider the items in I as team members and $g(I)$ as the total score. Given an itemset I and an item $i \in I$, the contribution $s_g(i, I)$ of i to $g(I)$ is:

$$s_g(i, I) = \sum_{J \subset I \setminus \{i\}} \frac{|J|!(|I| - |J| - 1)!}{|I|!} \left[g(J \cup i) - g(J) \right] \ . \tag{3}$$

The Shapley value $s_g(i, I)$ of i in I captures the notion of how much i contributed to the divergence or gap of I, and we have $\sum_{i \in I} s_g(i, I) = g(I)$. The Shapley values for $g(I)$ represent the local contribution of the items of the individual subgroup I to $g(I)$. The higher the value $s_g(i, I)$, the more the item i locally contributes to the total value $g(I)$.

Notation	*Description*
Item	Attribute equality with the form $a = v$, for an attribute $a \in A$ and a value v
Itemset (Subgroup) I	A set of items, each item referring to a distinct attribute
$\Delta_f(I, M)$	Divergence of Itemset I with respect to model M and statistic f
$gap_f(I, M_1, M_2)$	Cross-model performance gap from model M_1 to model M_2 for Itemset I and statistic f
$s_g(i, I)$	Shapley value of item i in I, showing how much i contributed to the divergence or gap of I $g(I)$
$\tilde{S}_g(i, u)$	Global Shapley value of item i, measuring the average effect to g of adding i to all other compatible itemset

TABLE I: Summary of the notation used in this work.

Global contribution. We evaluate the divergence or gap for all itemsets with adequate representation in the dataset, given by a frequency threshold u. The global contribution estimates the average role of an item on the divergence or gap, considering its effect on all explored itemsets. We consider the *global* Shapley value $\tilde{S}_g(i)$ of an item i, which measures the average effect of adding item i to all other compatible itemsets [15].

Let D be a dataset with attributes A, and let g be the gap or divergence of its itemsets measured for a given outcome function. Let $\mathcal{I}^\star_B$ be the set of frequent itemsets with attributes B for support threshold u. The *global divergence* $\tilde{S}_g(i, u)$ of a frequent item i is computed as follows:

$$\tilde{S}_g(i, u) = \tag{4}$$
$$\sum_{B \subseteq A \setminus attr(i)} \frac{|B|!(|A| - |B| - 1)!}{|A|! \prod_{b \in B \cup attr(i)} m_b} \sum_{J : J \cup i \in \mathcal{I}^\star_{B \cup attr(i)}} \left[g(J \cup i) - g(J) \right],$$

where $a = attr(i)$ is the attribute of item i. The global Shapley value $\tilde{S}_g(i, u)$ of i appropriately averages the effect of adding i to all itemsets not containing items for a. The computation accounts only for frequent itemsets to reduce the enumeration while ensuring the statistical significance of measure g over itemset. Further details are given in [15]. The advantage of using $\tilde{S}_g(i, u)$, rather than simply $g(i)$, is that it captures the incremental effect of adding item i to all other itemsets, rather than just measuring the effect on item i. We will use $\tilde{S}_g(i)$ when u is clear from the context to ease the notation.

Table I summarizes the concepts introduced in this section with the corresponding notation.

D. DivExplorer for subgroup exploration

We leverage DivExplorer [15] to extract itemsets and compute the statistic f and subgroup divergence or gap.

DivExplorer extracts all itemsets above a given support threshold, i.e., frequent ones. The support threshold u (such as 0.1% of the dataset) binds the exploration and ensures subgroups' statistical and operational significance. By completely slicing the metadata domain, the number of subgroups in the number of attributes is exponential. However, multiple

Dataset	*Samples*	*Subgroups*	*Avg Time*	*Worst Time*
FSC [20]	3793	47736	1.33s	1.40s
SLURP [21]	13078	3896	0.75s	0.81s
IEMOCAP [22]	4490	7932	1.03s	1.09s
LIBRISPEECH [16]	2620	2414	0.14s	0.19s

TABLE II: Average (across ten runs) and worst execution time [s] of DIVEXPLORER subgroup exploration. Note that FSC counts 47736 subgroups due to the high number of metadata.

extracted itemsets may have very small or empty support. These itemsets are less of interest for our subgroup performance analysis. Performance measures for subgroups with small support can be subject to statistical fluctuations that render divergence or gap measures not statistically significant. Itemsets that contain sufficient data are also operationally significant. Divergence or gap affecting a more significant portion of the dataset is more consequential than the ones that only affect a smaller amount. These reasons also suit our context. We will only consider frequent itemsets, that is, itemsets whose support size is above a given threshold u.

Finding frequent itemsets in a dataset is a fundamental task in data mining, known as frequent pattern mining [25], [26]. DIVEXPLORER augments frequent pattern mining techniques to efficiently extract frequent itemsets while computing the statistics f and divergence.

Given a dataset D with metadata attributes A, we leverage DIVEXPLORER to analyze a model M at the subgroup level with respect to statistic f. The result is the set of frequent itemsets over all the input metadata. For each frequent itemset I, we have its statistic $f(I, M)$, its divergence $\Delta_f(I, M)$, and its statistical significance t. To analyze the gap in performance between two models M_1 and M_2, we can adopt DIVEXPLORER separately for them. We can then directly compute for each frequent itemsets the gap in performance as the difference between $f(I, M_2)$ and $f(I, M_1)$.

In the next section, we show that the proposed methodology can reveal how model performance varies across subgroups and compare models at the subgroup level, identifying modeling biases towards peculiar subgroups.

IV. EXPERIMENTS AND RESULTS

We evaluate the performance of our approach by showing its ability to reveal sources of error (Section IV-A), analyzing how model size (Section IV-B1), architecture (Section IV-B2) and pre-training objective (Section IV-B3) impact performance at the subgroup level. Section IV-A also evaluates the difference between our analysis and a baseline approach. The experimental results show that our approach allows understanding model behaviors at the subgroup and global levels. They also highlight the generalizability of the method for multiple performance measures and tasks. Section IV-C discusses how our findings translate into practical insights.

We run the experiments on a machine equipped with Intel® Core™ i9-10980XE CPU, 2 × Nvidia® RTX A6000 GPU, 128 GB of RAM running Ubuntu 22.04 LTS. Once model fine-tuning and inference are performed, the subgroup exploration typically takes a few seconds. Table II summarizes the average and worst-case time for subgroup exploration with our approach on each dataset. We set the user-defined minimum support threshold u equal to 0.03 to ensure that all subgroups in our datasets are well-represented. For the smallest test set, LIBRISPEECH, the smallest subgroups will include at least 75 instances. This cardinality aligns with the standard practice requiring between 50 to 100 instances for reliable results [27]. We used Welch's t-statistic, denoted with t, to assess the statistical significance of performance gaps. Adopting a common rule of thumb [28], when Welch's t-statistic for a subgroup performance gap was larger than 2, we rejected the null hypothesis and identified the performance gap as statistically significant.

Datasets. We evaluated our approach on four datasets and three tasks. Specifically, we considered LIBRISPEECH [16] for the ASR task, FSC [20] and SLURP [21] for the IC task, IEMOCAP [22] for the ER task. We provided a detailed description of the datasets in Section V-A of the supplementary material. We considered as our target statistics the accuracy metric for FSC, SLURP, and IEMOCAP, while we adopted the Word Error Rate (WER) for LIBRISPEECH.

Metadata. We annotated the datasets with several metadata.

Speaker Demographics: We considered demographic metadata characterizing the speaker, if available. We examined all available (self-declared) demographic metadata within the datasets. Specifically, we used the gender, age, and country for FSC, the gender and country for SLURP, and the gender for LIBRISPEECH and IEMOCAP.

Speech-oriented metadata: We analyzed the number and the duration of silence, total and trimmed, the number of words, and the "speaking rate" as the number of words per second. Note that trimmed duration stands for the duration of the utterance without considering the first and the last pause. In contrast, "total silence" duration denotes the duration of an utterance without considering any pause. We observed that the number and length of intermediate pauses did not significantly affect the performance of the models across the various datasets and tasks, except for LIBRISPEECH. Consequently, we chose to retain them solely for this dataset, given their important role in achieving accurate ASR.

Dataset-dependent metadata: We examined the metadata specific for each dataset and/or task, if available. We considered the intent for FSC and SLURP as the combination of action, object, and location for the former, action and scenario for the latter, and both the categorical-based and attribute-based labels for IEMOCAP.

Table III summarizes the metadata we collected and analyzed for the considered datasets, separately for each domain. These metadata cover different aspects of speech data, ranging from demographics to speaking conditions. The flexibility of our approach allows practitioners to expressly define and adopt any metadata tailored to specific application contexts.

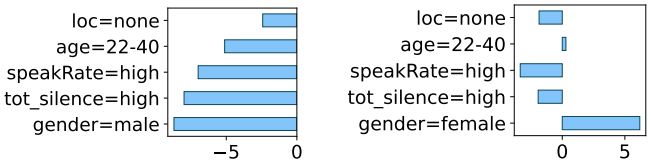
Metadata discretization. Attributes that are continuous in nature, such as speaking rate or utterance duration, need to be

Task	*Dataset*	*Demographics Metadata*	*Speech-Related Metadata*	*Dataset-dependent Metadata*
IC	FSC	gender, age, country	number and duration of silences, speech rate, number of words	intent (action, object, location)
IC	SLURP	gender, country	number and duration of silences, speech rate, number of words, close/far field	intent (action, scenario)
ER	IEMOCAP (IEMO)	gender	number and duration of silences, speech rate, number of words	emotion label arousal labels (activation, valence, dominance)
ASR	LIBRISPEECH (LS)	gender	number and duration of silences, speech rate, number of words, number and duration of middle pauses	none

TABLE III: Summary of each dataset collected demographic, speech-related, and dataset-dependent metadata.

Task	*Dataset*	w2v2-b	w2v2-l	hub-b	hub-l
IC	FSC	91.72	93.17	98.42	**98.50**
IC	SLURP	86.86	85.59	87.69	**89.25**
ER	IEMOCAP	74.66	71.18	67.44	**74.99**
ASR	LIBRISPEECH	6.06	3.82	6.56	**3.50**

TABLE IV: Overall performance on the selected speech-related datasets. We report Accuracy (%) for IC and ER tasks and WER (%) for ASR. Best results are in boldface.



(a) w2v2-b. $\Delta_{acc} = -31.22\%$ (b) w2v2-b. $\Delta_{acc} = -0.65\%$

Fig. 1: RQ1. FSC DATASET. Item contribution to accuracy for (a) the subgroup with the highest negative divergence (*Sup*=0.03) and (b) when considering female gender (*Sup*=0.04) rather than male.

discretized into fixed ranges. In this work, we discretized these metadata in three ranges using frequency-based discretization, and we renamed the ranges as "low," "medium," and "high."

Models. We considered the mono-lingual wav2vec 2.0 [17] and HuBERT [23] models of two different sizes, base (ca. 90 million parameters) and large (ca. 300 million parameters). For FSC and IEMOCAP, we used the public fine-tuned checkpoints [29], while for SLURP and LIBRISPEECH, we followed fine-tuning procedures and guidelines from relevant literature [30]. Table IV highlights the performance of the fine-tuned models on each of these datasets. Additionally, for the impact of the mono- and multi-lingual pre-training objective on the subgroup performance (Section IV-B3), we leveraged XLSR large (with 300 million parameters) models pre-trained on 53 [31] and 128 [32] languages.

A. Model understanding at the subgroup level: identification of the most problematic intra-model subgroups

Here we address **RQ1**. We apply our methodology to detect subgroups that diverge from the average behavior. We recall that our approach involves exploring all adequately represented subgroups across metadata and assessing their divergence from the average behavior. A subgroup is deemed problematic if its divergence is large in value, and statistically significant (as determined via the Welch t-test). The specific threshold for divergence to be deemed of large value depends on the nature of the problem under analysis; what may be acceptable in one context may not be in another. In the following results, we consider the wav2vec 2.0 base model and all datasets in the analysis. **Table V** highlights the most negatively and positively divergent (i.e., problematic) subgroups for each dataset. The divergence values for these subgroups are all significant ($t > 2$, following [28]'s rule of thumb).

FSC. We examine model accuracy in data subgroups. The higher the accuracy, the better the results. Thus, a negative divergence denotes an accuracy lower than average. Conversely, a positive divergence indicates a higher one.

wav2vec 2.0 base achieves the worst performance for the subgroup {*"age=22-40, gender=male, location=none, speaking rate=high, tot silence=high"*}, reported in the first block of Table V, with divergence $\Delta_{acc} = -31.2\%$. Analyzing the influence of sensitive attributes, such as gender, is particularly relevant. If we consider the female gender for this subgroup while keeping the other metadata values constant, the subgroup performance rises. Hence, for the identified subgroup, female speakers achieve better accuracy than males. The Shapley values in Figure 1(a) also confirm that the male gender is associated with lower accuracy. Conversely, the female gender has a positive impact and leads to higher accuracy scores (Figure 1(b)).

Divergence analysis also reveals subgroups with better performance than average. The most positively divergent subgroup consists of utterances of speakers aged 22-40 with a low speaking rate and long duration, having "washroom" as the target location.The model correctly predicts all utterances in this subgroup.

We can assess the influence of each metadata value on divergence using the global Shapley values, as depicted in Figure 2. Metadata values with negative global Shapley value identify population characteristics that, when added

Dataset	*Subgroups*	*Sup_train*	*Sup_test*	f	Δ_f	t
FSC	S^-: {*"age=22-40, gender=male, location=none, speaking rate=high, tot silence=high"*}	0.03	0.04	60.50	-31.22	7.05
	S^+: {*"age=22-40, location=washroom, speaking rate=low, trimmed duration=high"*}	0.03	0.03	100.0	8.28	9.74
SLURP	S^-: {*"action=quirky"*}	0.04	0.05	67.37	-19.50	10.27
	S^+: {*"gender=female, scenario=weather"*}	0.03	0.03	95.93	9.07	8.32
IEMO	S^-: {*"label=happy, activation=low"*}	0.03	0.03	44.74	-29.92	7.37
	S^+: {*"label=sad, valence=low, tot silence=low, trimmed duration=high"*}	0.03	0.03	98.57	23.92	17.01
LS	S^-: {*"gender=female, trimmed speaking rate=high, trimmed duration=low, num pauses=low"*}	0.05	0.03	17.30	11.24	4.16
	S^+: {*gender=female, speaking rate=low, trimmed speaking rate=low, num pauses=low, tot duration=medium*}	0.03	0.03	3.27	-2.79	5.57

TABLE V: RQ1. Gap in performance measure f (accuracy for FSC, SLURP, IEMOCAP, WER for LIBRISPEECH) for the most negatively (S^-) and positively (S^+) divergent subgroups compared to overall test performance; wav2vec 2.0 base. The t column indicates the Welch's t-test value.

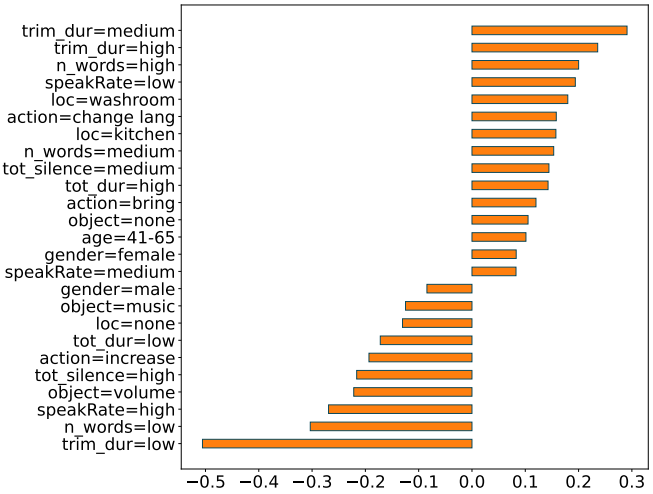


Fig. 2: RQ1. Global Shapley values of Accuracy divergence; FSC dataset, wav2vec 2.0 base model. Terms showing negative contributions indicate a lower-than-average accuracy.

to subgroups, lower accuracy on average, while metadata with positive global Shapley values rise accuracy on average. Speaking conditions notably influence performance: shorter durations, fewer words, and faster speaking rates have negative global Shapley value, whereas longer durations, more words, and slower speaking rates have positive values. These speaking conditions align with factors highlighted in [33] that influence error rates, confirming the observed subgroup behaviors. Furthermore, intent targets have a distinct impact: the object "volume" negatively impacts performance, whereas the "washroom" location is associated with higher accuracy.

SLURP. The subgroup {*"action = quirky"*} experiences the highest drop in accuracy. Hence, for this dataset and the wav2vec 2.0 model, the target action equal "quirky" *alone* is the term with the highest error rate, with a divergence Δ_{acc} equal to -19.5% (second block of Table V).

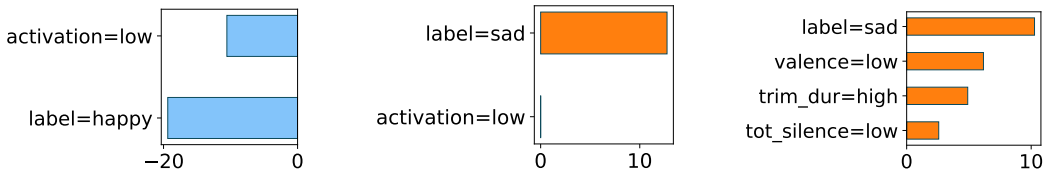
IEMOCAP is a dataset for the ER task, where utterances are annotated with emotion labels. We analyze performance variation for the class labels in *conjunction* with other emotions or speaking conditions.

The analysis of the most divergent subgroups reveals that happiness is associated with lower-than-average performance, especially in conjunction with other attributes. The subgroup {*"label=happy, activation=low"*} has the lowest performance (third block of Table V). Note that the emotion {*"label=happy"*} alone achieves an accuracy of 64.03%. This highlights the relevance of its association with low activation, showing a significant performance drop. To understand each item importance, we can inspect this subgroup local Shapley values, i.e., analyze each item contribution to the subgroup performance. The Shapley values depicted in Figure 3(a) shows the high impact of the label "happy" followed by the "low activation". If we change the label to "sad", the performance hugely increases (87.41%) (Figure 3(b) shows the local Shapley values of the latter subgroup).

On the other hand, the label "sad", in conjunction with other items, is associated with the most positively divergent subgroups, with accuracy higher than average. Figure 3(c) shows the impact of each subgroup item. We notice the predominant role of the sad label compared to the others.

LIBRISPEECH. We analyze the Word Error Rate (WER) in data subgroups. A subgroup's inferior performance relative to the overall system is indicated by a higher divergence of its WER value. Hence, unlike the previous cases, a positive WER divergence indicates lower performance.

The subgroup with the highest positive divergence is {*"gender=female, trimmed speaking rate=high, trimmed duration=low, num pauses=low"*}. Hence, the model exhibits a higher error rate for short utterances, low pauses, high speaking rates, and the female gender. We now analyze the impact of gender. When considering the male gender while keeping other speaking conditions fixed, the models perform better, with a WER score of 9.89% for wav2vec 2.0 base. This indicates a disparate impact of gender on performance for this subgroup. The Shapley values (Figure 4) confirm the positive influence of the male gender, associated with lower WER. This aligns with existing literature on the ASR and Speaker Recognition tasks ([34], [35], [36], [37], [38], [39], [40], [41]), in which male speakers and speakers with lower speaking rates are associated with higher performance. Conversely, female speakers and faster-speaking speakers exhibit lower performance. We report the impact of each item on the WER divergence using the global Shapley values in the Supplementary material (Section



(a) Most negatively divergent subgroup. acc=44.74%, sup=0.03.

(b) Same subgroup as (a), with "sad" label. acc=87.41%, sup=0.15.

(c) Most positively divergent subgroup. acc=98.57%, sup=0.03.

Fig. 3: RQ1. Item contribution to performance. IEMOCAP dataset, wav2vec 2.0 base.

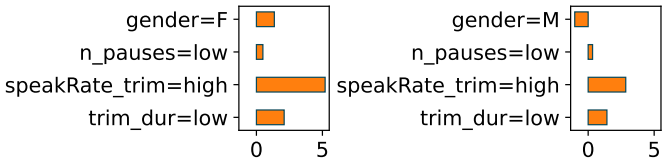


Fig. 4: RQ1. Item contribution to performance within the same subgroup, either considering *gender=female* (left; WER 17.30%, support=0.03) or *gender=male* (right; WER 9.89%, support=0.04). LIBRISPEECH; wav2vec 2.0 base.

V-B).

Comparison with a baseline. Traditional subgroup analysis often examines performance disparities based on a single attribute, such as gender or speaking rate. However, considering subgroups at the intersection of multiple attributes allows for a more comprehensive understanding of performance dynamics. By exploring these intersectional subgroups, we can uncover intricate relationships and patterns that may remain hidden when analyzing subgroups based on one single attribute.

We compare the highest and lowest divergence scores identified by our method (the ones reported in Table V) with the ones obtained when only one-level subgroups are considered. We report the comparison results in Figure 5. Our approach demonstrates superior performance over the baseline. For FSC, IEMOCAP, and LIBRISPEECH, the gap in identified divergence is extremely large. For SLURP, the two approaches exhibit comparable outcomes, because the subgroups identified by our approach mostly coincide with one-level subgroups. Our approach surpasses the baseline by successfully identifying subgroups that exhibit greater divergence than simpler one-level subgroups. This observation also holds for all other models and sizes considered in our analysis (wav2vec 2.0 large, HuBERT base, and large), as reported in the supplementary material.

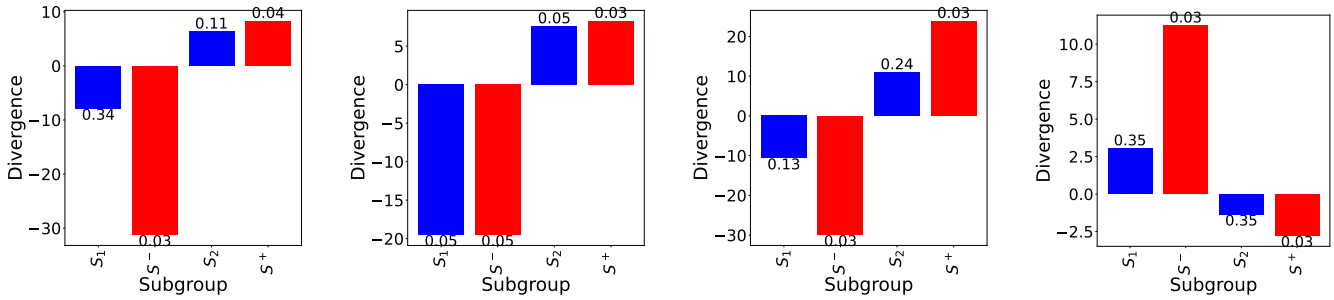
Summary of findings. We identified subgroups that exhibit substantial deviations from the average performance of the model, along with the corresponding error sources. By examining the influence of metadata values on divergence at both local and global levels, we gained insights for model debugging and comprehension. Our study uncovered the disparate impact of gender on performance, illustrating the utility of our approach as a tool for fairness evaluation. Moreover, we showcased that the simultaneous consideration of multiple metadata values, rather than isolated factors, enabled the detection of highly divergent behaviors. These findings show the proposed approach effectiveness in identifying subgroups with pronounced divergence and highlight its potential to enhance the performance of the analyzed models.

B. Model Comparisons

Identifying effective machine learning models capable of achieving superior performance across diverse subgroups represents a critical challenge in contemporary data-driven research. In this section, we present a comprehensive analysis that aims to compare the performance of different models at the overall and subgroup levels to examine which subgroups are most likely to benefit from model modifications and which are not. Our approach is entirely model-agnostic, enabling us to extend our analysis to compare models with different sizes (Section IV-B1), with entirely dissimilar architectures (Section IV-B2) or trained with distinct pre-training objectives (mono vs. multilingual, Section IV-B3). Through this comparative analysis, we can identify models that offer the most significant potential for enhancing subgroup performance separately per dataset and task, thereby paving the way for more equitable and inclusive data-driven research outcomes.

1) Effect of the model size on subgroup performance: Here, we address **RQ2**. Larger machine learning models are generally more accurate than smaller ones. [19] claims that larger models are also fairer. However, we recently demonstrated [24] that increasing the size of a model does not always lead to better performance on a given dataset, as there may be subgroups within the data for which the model's performance decreases. In [24], we thoroughly investigated the FSC dataset and the effect of scaling size. To augment the empirical evidence for this phenomenon, we conducted experimental evaluations on the other three datasets. **Table VI** summarizes the performance gap when scaling up the model size of wav2vec 2.0 for each dataset, highlighting the subgroups with the highest performance improvement and the highest decrease.

FSC. Scaling up wav2vec 2.0 for FSC offers advantages on a broader scale, encompassing both overall performance improvements (93.17% vs. 91.72% accuracy) and specific subgroup enhancements. Our analysis reveals that in 63.75% of the examined subgroups, the larger model yields performance improvements. Notably, the subgroup {*action=increase, location=none, tot duration=low, trimmed speaking rate=low,*

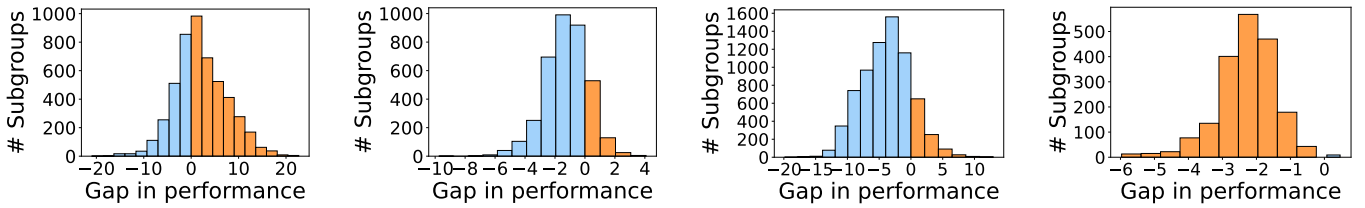


(a) FSC dataset. (b) SLURP dataset. (c) IEMOCAP dataset. (d) LIBRISPEECH dataset.

Fig. 5: Comparison with a baseline considering the most divergent one-level subgroups. Blue color indicates the baseline performance (S_1 and S_2), while red denotes our approach (S^- and S^+). wav2vec 2.0 base model. The numbers above and below the bars indicate the support of each considered subgroup.

Dataset		*Subgroups*	*Sup*	gap_f	$f_{\texttt{w2v2-b}}$	$f_{\texttt{w2v2-l}}$	t
FSC	↑	{*action=increase, location=none, tot duration=low, trimmed speaking rate=low, trimmed duration=low*}	0.03	22.69	75.63	98.32	5.37
	↓	{*"action=activate, gender=male, speaking rate=low"*}	0.03	-20.97	96.77	75.81	4.92
SLURP	↑	{*gender=female, speaking rate=high, trimmed speaking rate=high, trimmed duration=low*}	0.04	4.08	83.88	87.96	1.83
	↓	{*"action=remove, num words=low"*}	0.03	-9.74	92.64	82.90	4.33
IEMO	↑	{*"label=happy, trimmed speaking rate=low"*}	0.04	12.96	67.28	80.25	2.66
	↓	{*"label=sad, trimmed speaking rate=low"*}	0.03	-19.86	70.55	50.68	3.53
LS	↑	{*gender=female, num pauses=low, trimmed speaking rate=high, trimmed duration=low*}	0.03	-5.97	17.30	11.33	1.78
	↓	{*gender=male, num pauses=low, tot duration=low, trimmed speaking rate=high, trimmed duration=low*}	0.04	0.46	10.17	10.64	0.14

TABLE VI: RQ2. Performance gap for performance measure f (WER for LIBRISPEECH, accuracy for the others) when scaling up wav2vec 2.0 size, from base (90 million parameters) to large (300 million parameters). (↑) denotes the highest performance improvement, (↓) indicates the largest decrease. The t column indicates the the Welch's t-test value.



(a) FSC. Performance improvement for 63.75% of subgroups, decrease for 31.89% of them.

(b) SLURP. Performance improvement for 16.86% of subgroups, decrease for 80.11% of them.

(c) IEMOCAP. Performance improvement for 11.21% of subgroups, decrease for 85.85% of them.

(d) LIBRISPEECH. Performance improvement for 99.25% of subgroups, decrease for 0.75% of them.

Fig. 6: RQ2. Gap contribution when scaling up wav2vec 2.0, considering different datasets.

trimmed duration=low"} exhibits the most significant enhancement, with a remarkable improvement of 22.69%, as illustrated in the first section of Table VI (t=5.37). However, in 31.89% of the subgroups, performance decreases, with the most substantial negative impact observed in the subgroup {*action=activate, gender=male, speaking rate=low"*}, resulting in a maximum decline of 20.97%.

SLURP. Similar overall performance does not entail comparable performance also at the subgroup level. The base and large wav2vec 2.0 models perform similarly (86.86% for base and 85.59% for large). However, wav2vec 2.0 base outperforms the large on 80.11% of the explored subgroups. Table VI presents the subgroups with the highest decrease (−9.74%) and the highest improvement (4.08%) when scaling up this model. Figure 6(b) shows the gap distribution, which is overall balanced with a peak at around −1%.

IEMOCAP. Scaling wav2vec 2.0 for IEMOCAP is not beneficial overall and at the subgroup level. The accuracy drops from 74.66% of the base version to 71.18% of the large one. For many explored subgroups (85.85%), performance *decreases* when scaling up wav2vec 2.0. Table VI reports the subgroups with the highest decrease ({*"label=sad, trimmed speaking rate=low"*}, with a negative gap of −19.86%) and with the largest improvement ({*"label=happy, trimmed speaking rate=low"*}, with a positive gap of 12.96%) when scaling up the model. The gap distribution, reported in Figure 6(c), is

skewed towards negative values, peaking at around -3%.

LIBRISPEECH. The base version of wav2vec 2.0 model achieves an overall WER score of 6.06%, while the large a much lower (thus, better) 3.82%. The improvement is observed for almost all explored subgroups (99.25%). Hence, the larger model shows *better performance both overall and at the subgroup level*. The highest improvement is -5.97% for the subgroup {*"gender=female, num pauses=low, trimmed speaking rate=high, trimmed duration=low"*} (Table VI). Yet, while we observe a slight improvement, the performance gap is not statistically significant. Figure 6(d) shows the gap distribution when scaling up the wav2vec 2.0 model, with peak at around -2.5%.

Summary of findings. For the LIBRISPEECH dataset, we demonstrate that scaling up the model enhances both overall and subgroup performance. However, in the IEMOCAP dataset, enlarging the model leads to lower performance overall and within subgroups. Our evaluations on the FSC dataset suggest that the efficacy of a larger model might vary significantly based on the specific subgroup under examination. Furthermore, our studies with the SLURP dataset highlight that achieving similar overall performance does not necessarily translate to similar performance also at the subgroup level.

Our findings suggest that the relationship between model size and performance is complex, dataset- and task-dependent. While increasing model size may lead to better performance in some cases, it is essential to consider subgroup performance carefully when evaluating larger models' effectiveness.

2) Effect of the model architecture on the subgroup performance: Here, we address **RQ3**. We adopt the proposed methodology to evaluate the performance enhancement obtained by replacing a particular model architecture with a different one. Specifically, we evaluate the performance gaps obtained by replacing the wav2vec 2.0 base model with the HuBERT base model. We show that even when the overall performance increases when one model is adopted instead of another, there can be subgroups where performance decreases. Similar considerations apply to the analysis of wav2vec 2.0 to HuBERT larger models, reported in the supplementary material.

Table VII outlines the performance gap when changing the models' architecture, highlighting the subgroups with the highest performance improvement and the highest decrease.

FSC. HuBERT base outperforms wav2vec 2.0 base (98.42% vs. 91.72%). Upgrading to HuBERT results in a positive gap for most subgroups (97.03% for the base models), with a peak of approximately 5% accuracy gap observed in the distribution (Figure 7(a)). The first block of Table VII reports the subgroups with the most considerable improvement and decrease in accuracy when changing the model structure.

SLURP. HuBERT outperforms wav2vec 2.0 in overall accuracy. At the subgroup level, the HuBERT base model performs better than wav2vec 2.0 in most explored subgroups (77.16%). The distribution of accuracy gap is shown in Figure 7(b). The maximum gap is by 5.46% for the subgroup {*"field=far, gender=male, tot duration=high, tot silence=low, trimmed duration=high"*}, as shown in the second block of Table VII (t=2.13). Yet, 16.86% of explored subgroups experience a decrease in performance, with a drop that goes up to -9.74%.

IEMOCAP. At a global level, the performance of HuBERT base is lower than wav2vec 2.0 base (67.44% vs. 74.66%). Regarding subgroup analysis, 93.95% of the examined subgroups did not exhibit improvement upon transitioning from wav2vec 2.0 to HuBERT base. The highest negative impact on accuracy (-30.14%) was observed for the subgroup {*"label=sad, trimmed speaking rate=low"*} (third block of Table VII, t=5.41.

LIBRISPEECH. The wav2vec 2.0 base has a slightly better performance than the HuBERT base, with a WER of 6.06% and 6.56%, respectively. Switching from the wav2vec 2.0 base to the HuBERT base introduces a decline in performance for 89.77% of the analyzed subgroups. Still, the higher increase in WER, observed for the subgroup {*"gender=male, num pauses=low, num words=low, tot silence=medium"*} (last block of Table VII), is not statistically significant ($t = 1.33$). The distribution of the WER gap shows a peak around 0.5% (Figure 7(d)).

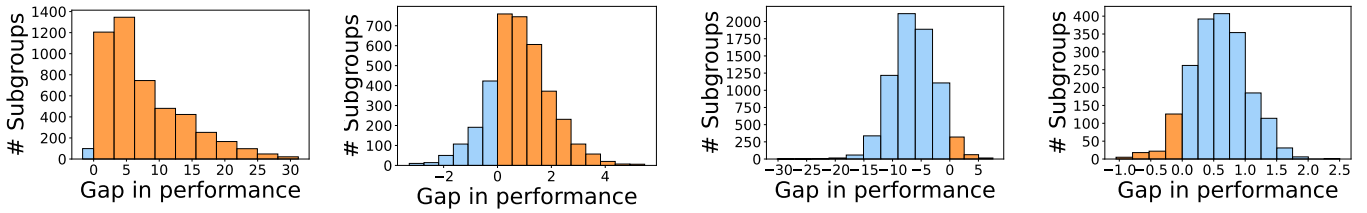
Summary of findings. We show that changing architecture from wav2vec base to HuBERT base benefits almost all subgroups for FSC and most subgroups for SLURP. On the other hand, the change harms most subgroups for IEMOCAP and LIBRISPEECH. Hence, the results show the disparate impact of model architectures and the complexity of the relationship between architecture and performance. Our findings show the limitation of overall model comparison and motivate the need to compare subgroup performance across architectures.

3) Comparison between multi- and mono-lingual pretrained models: Here, we address **RQ4**. We investigate the impact of the multi-lingual pre-training objective on model behavior. Specifically, we compare the performance of three large models, namely the mono-lingual wav2vec 2.0 [17] model and the multi-lingual XLSR-53 [31] and XLSR-128 [32] models, on the FSC dataset. We select this dataset for its widespread use in literature, modest size, and previous examination by researchers at the subgroup level [24]. **Table VIII** show the highest and lowest performance gap when transitioning the models' pre-training objective from mono-lingual to multi-lingual. Figure 8 shows the gap distribution when changing the pre-training objective from mono- to multi-languages. Specifically, Figure 8(a) presents the gap distribution when transitioning to XLSR-53, while Figure 8(b) from mono to XLSR-128. We observe different trends. The former displays a left-skewed distribution, with a predominance of negative values, whereas the latter reveals a right-skewed distribution, mainly characterized by positive values.

Mono- to Multi- Languages (XLSR-53). The XLSR-53 large model performs less effectively than the large mono-lingual: the former achieves an accuracy of 90.07% while the latter 93.17%. The XLSR-53 model is worse at the subgroup level than the mono-lingual on 72.30% of explored subgroups. The decrease goes up to 55.83% ($t = 11.49$). Only 25.59% of the

Dataset		*Subgroups*	*Sup*	gap_f	$f_{\texttt{w2v2-b}}$	$f_{\texttt{hub-b}}$	t
FSC	↑	{*"gender=male, location=none, num words=low, tot silence=high, trimmed duration=low"*}	0.03	31.20	64.00	95.20	6.53
	↓	{*"action=decrease, age=22-40, location=washroom"*}	0.03	-1.68	100.00	98.32	1.01
SLURP	↑	{*"field=far, gender=male, tot duration=high, tot silence=low, trimmed duration=high"*}	0.03	5.46	80.76	86.22	2.13
	↓	{*"field=far, gender=female, speaking rate=low, tot duration=low, tot silence=low"*}	0.04	-3.27	85.81	82.53	1.35
IEMO	↑	{*"activation=high, label=anger, duration=low, valence=low"*}	0.03	7.54	75.34	82.88	1.57
	↓	{*"label=sad, trimmed speaking rate=low"*}	0.03	-30.14	70.55	40.41	5.41
LS	↑	{*"num pauses=medium, speaking rate=medium, tot duration=medium, tot silence=medium"*}	0.04	-1.05	7.44	6.39	0.75
	↓	{*"gender=male, num pauses=low, num words=low, tot silence=medium"*}	0.03	2.5	7.60	10.11	1.33

TABLE VII: RQ3. Gap for performance measure f when changing the models' architecture, from wav2vec 2.0 to HuBERT base. (↑) denotes the highest performance improvement, (↓) indicates the largest decrease. The t column indicates the Welch's t-test value.



(a) FSC. Performance improvement for 97.03% of subgroups, decrease for 1.04% of them.

(b) SLURP. Performance improvement for 77.16% of subgroups, decrease for 18.99% of them.

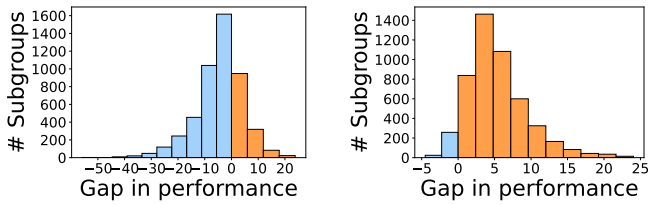
(c) IEMOCAP. Performance improvement for 4.15% of subgroups, decrease for 93.95% of them.

(d) LIBRISPEECH. Performance improvement for 10.23% of subgroups, decrease for 89.77% of them.

Fig. 7: RQ3. Gap distribution when changing wav2vec 2.0 base to HuBERT base.

Dataset		*Subgroups*	*Sup*	gap_f	$f_{\texttt{mono}}$	$f_{\texttt{xlsr-53}}$	t
FSC	↑	{*"action=activate, gender=male, trimmed speaking rate=low"*}	0.04	23.84	74.83	98.68	6.39
	↓	{*"action=increase, object=heat, trimmed speaking rate=medium, trimmed duration=high"*}	0.03	-55.83	96.67	40.83	11.49
		Subgroups	*Sup*	gap_f	$f_{\texttt{mono}}$	$f_{\texttt{xlsr-128}}$	*t-test*
	↑	{*"gender=male, speaking rate=high, tot silence=high, trimmed duration=low, location=none"*}	0.03	24.06	75.19	99.25	6.14
	↓	{*"action=increase, object=heat, age=22-40, gender=male, tot silence=low"*}	0.04	-4.51	98.50	93.98	1.79

TABLE VIII: RQ4. Performance gap when changing the pre-training objective from mono- to multi-lingual (XLSR-53 and XLSR-128m, respectively), FSC dataset. (↑) denotes the highest performance improvement, (↓) indicates the largest decrease. The t column indicates the Welch's t-test value.



(a) Mono- to XLSR-53. Performance improvement for 25.59% of subgroups, decrease for 72.30% of them.

(b) Mono- to XLSR-128. Performance increase for 92.90% of subgroups, decrease for 4.05% of them.

Fig. 8: RQ4. Gap distribution when changing the pre-training objective from mono- to XLSR-53 (a) and from mono- to XLSR-128 (b). FSC dataset.

subgroups experience a performance gap, with an increase up to 23.84% ($t = 6.39$).

Mono- to Multi- Languages (XLSR-128). Differently, switching to the XLSR-128 model is beneficial, *both overall and at the subgroup level*. The multi-lingual model achieves an overall 98.34% of accuracy (compared to 93% of the mono-lingual). At the subgroup level, it performs better than the mono-lingual on 92.90% of the explored subgroups. As shown in the second-last row of Table VIII, the most significant improvement is by 24.06% (t=6.14). We are interested in analyzing the 4.04% of subgroups for which we observe a decrease in performance instead. The highest decrease is small and not statistically significant (-4.5%, t=1.79) and the performance is still higher than the average of the mono-lingual model. This suggests the multi-lingual XLSR-128 model is more robust and suitable, considering the subgroup level performance.

Global item role in performance gap. We use the notion of global Shapley value to quantify the contribution of metadata

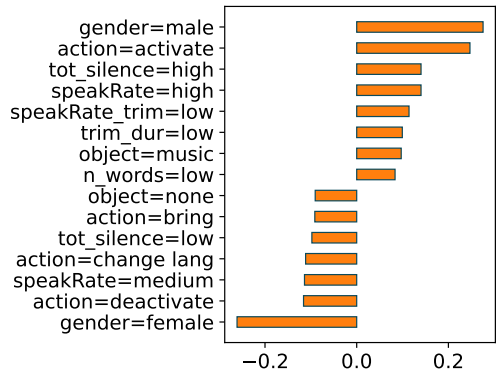


Fig. 9: RQ4. Global Shapley values of accuracy gap for wav2vec 2.0 large mono-lingual to XLSR-128, top-15. FSC.

values to a gap in performance when altering the pre-training objective. Figure 9 summarizes the top 15 items with the highest impact on the performance gap when changing the pre-training objective from mono-lingual to multi-lingual XLSR-128. The significant role of gender in determining the performance of the models is highlighted, with "*gender = female*" exerting a negative impact on the performance. In contrast, "*gender=male*" exhibits a positive influence on the performance gap. Moreover, a high speaking rate, a high silence, and specific actions (such as "activate") and objects (such as "music") have a positive impact when going from the mono-lingual model to XLSR-128. The findings thus underscore the importance of gender, speaking rate, and speech duration in designing and optimizing speech recognition models.

Summary of findings. Our study demonstrates that switching from the mono-lingual model to the XLSR-53 model for FSC does not provide any benefits in terms of performance, either at the overall or subgroup levels. Conversely, employing the XLSR-128 model has several advantages, as it outperforms the mono-lingual version overall and at the subgroup level. These findings highlight the need for thoroughly evaluating each model for the specific task and dataset at hand.

C. Final remarks

Our approach enables the analysis of model performance, and the comparison of models, at the subgroup level. Its adoption yields a series of practical insights. (i) By understanding which subgroups benefit or are disadvantaged by the adoption of a specific model, practitioners can determine if they can trust the model. (ii) Our approach facilitates model debugging by identifying subgroups affected by below-average performance. (iii) Practitioners can then actively work to improve the model. (iv) Finally, our approach aids practitioners in comparing models, allowing them to choose models based on subgroup-level performance criteria.

V. Conclusions

This study introduces a novel approach to evaluate spoken language understanding (SLU) system performance at the subgroup level, employing model bias analysis. Our methodology automates the detection of significant performance disparities in subgroups, facilitating error analysis and model comparison. The approach is applicable across various speech tasks, models, and metrics, making it widely generalizable. Through a comprehensive analysis of diverse datasets, tasks, and models, we demonstrate its effectiveness. The subgroup-level analysis provides a more nuanced assessment of model performance, allowing for identifying subgroups that benefit most from system improvements. Overall, this study advances the understanding of SLU system performance at the subgroup level and provides a valuable tool for developing more inclusive and effective speech technologies. In future work, we envision adapting the proposed methodology to model improvement.

Acknowledgment

This work was partially supported by the *Explaining Model Bias and Behavior for End-to-End Spoken Language Understanding (SLU) Models* collaboration with Amazon AGI. The work is then partially supported by the FAIR - Future Artificial Intelligence Research and received funding from 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) and the spoke "FutureHPC & BigData" of the ICSC - Centro Nazionale di Ricerca in High-Performance Computing, Big Data and Quantum Computing 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.

References

[1] R. Sarikaya, P. A. Crook, A. Marin, M. Jeong, J.-P. Robichaud, A. Celikyilmaz, Y.-B. Kim, A. Rochette, O. Z. Khan, X. Liu *et al.*, "An overview of end-to-end language understanding and dialog management for personal digital assistants," in *2016 ieee spoken language technology workshop (slt)*. IEEE, 2016, pp. 391–397.

[2] M. Nuruzzaman and O. K. Hussain, "A survey on chatbot implementation in customer service industry through deep neural networks," in *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*. IEEE, 2018, pp. 54–61.

[3] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, "Speech technology for healthcare: Opportunities, challenges, and state of the art," *IEEE Reviews in Biomedical Engineering*, 2020.

[4] G. Ioannides, M. Owen, A. Fletcher, V. Rozgic, and C. Wang, "Towards Paralinguistic-Only Speech Representations for End-to-End Speech Emotion Recognition," in *Proc. INTERSPEECH 2023*, 2023, pp. 1853–1857.

[5] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

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

[7] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proc. of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[8] Z. Mengesha, C. Heldreth, M. Lahav, J. Sublewski, and E. Tuennerman, ""i don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on african americans," *Frontiers in Artificial Intelligence*, p. 169, 2021.

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

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

[11] C. Liu, M. Picheny, L. Sarı, P. Chitkara, A. Xiao, X. Zhang, M. Chou, A. Alvarado, C. Hazirbas, and Y. Saraf, "Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6162–6166.

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

[13] L.-F. Lai and N. Holliday, "Exploring Sources of Racial Bias in Automatic Speech Recognition through the Lens of Rhythmic Variation," in *Proc. INTERSPEECH 2023*, 2023, pp. 1284–1288.

[14] E. Kim, Y. Chae, J. Sim, and K. Lee, "Debiased Automatic Speech Recognition for Dysarthric Speech via Sample Reweighting with Sample Affinity Test," in *Proc. INTERSPEECH 2023*, 2023, pp. 1508–1512.

[15] 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.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[17] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

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

[19] Y. Sheng, J. Yang, Y. Wu, K. Mao, Y. Shi, J. Hu, W. Jiang, and L. Yang, "The larger the fairer? small neural networks can achieve fairness for edge devices," *arXiv preprint arXiv:2202.11317*, 2022.

[20] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, 2019, pp. 814–818.

[21] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "SLURP: A spoken language understanding resource package," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7252–7262.

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

[23] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[24] 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.

[25] R. Agrawal, R. Srikant *et al.*, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215. Santiago, Chile, 1994, pp. 487–499.

[26] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '00, 2000, p. 1–12.

[27] J. Hair, W. Black, B. Babin, and R. Anderson, *Multivariate Data Analysis*. Cengage, 2019. [Online]. Available: https://books.google.it/books?id=0R9ZswEACAAJ

[28] A. F. Siegel, "Chapter 10 - hypothesis testing: Deciding between reality and coincidence," in *Practical Business Statistics (Sixth Edition)*, sixth edition ed., A. F. Siegel, Ed. Springer Science & Business Media, 2012, pp. 249–287.

[29] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," pp. 1194–1198, 2021.

[30] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[31] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[32] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[33] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, "ERASER: A benchmark to evaluate rationalized NLP models," in *ACL 2020*, 2020.

[34] C. S. Greenberg, D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki, and D. A. Reynolds, "The nist 2014 speaker recognition i-vector machine learning challenge," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014, pp. 224–230.

[35] M. Senoussaoui, P. Kenny, N. Brümmer, E. d. Villiers, and P. Dumouchel, "Mixture of plda models in i-vector space for gender-independent speaker recognition," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[36] S. Cumani, N. Brümmer, L. Burget, P. Laface, O. Plchot, and V. Vasilakakis, "Pairwise discriminative speaker verification in the i-vector space," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1217–1227, 2013.

[37] S. Cumani, O. Glembek, N. Brümmer, E. De Villiers, and P. Laface, "Gender independent discriminative speaker recognition in i-vector space," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4361–4364.

[38] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and rsr2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[39] R. González Hautamäki, V. Hautamäki, and T. Kinnunen, "On the limits of automatic speaker verification: Explaining degraded recognizer scores through acoustic changes resulting from voice disguise," *The Journal of the Acoustical Society of America*, vol. 146, no. 1, pp. 693–704, 2019.

[40] S. M. Chu and D. Povey, "Speaking rate adaptation using continuous frame rate normalization," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4306–4309.

[41] F. Tong, S. Zheng, H. Zhou, X. Xie, Q. Hong, and L. Li, "Deep representation decomposition for rate-invariant speaker verification," *arXiv preprint arXiv:2205.14294*, 2022.

SUPPLEMENTARY MATERIAL

This document presents additional results that provide a comprehensive analysis across the whole set of datasets, tasks, and models of the four research questions we addressed in our work.

We include, for completeness, a rich collection of figures and tables that provide visual representations and qualitative measurements to further support our main findings.

A. Dataset characterization

LIBRISPEECH [16] corpus is a collection of audio recordings sourced from audio books belonging to the LibriVox initiative. It encompasses a corpus of 1000 hours of speech sampled at a rate of 16 kHz. For our experiments, we used the *"clean-100"* version, which comprises 100 hours of clean audio samples. The test set is characterized by 2620 samples recorded by 40 different speakers. The evaluation metric for the ASR task is the Word Error Rate (WER).

FLUENT SPEECH COMMANDS (FSC) [20] is a dataset widely employed for the Intent Classification (IC) task. The test set of FSC consists of 3793 audio samples mapped to 31 unique intents and has been recorded by ten speakers. Each audio sample corresponds to three slots: action, object, and location. The combination of the aforementioned slots determines the intent of each audio sample. The IC task is evaluated using intent accuracy as the metric.

SLURP [21] dataset is a collection of audio recordings designed for audio Intent Classification. It consists of audio samples recorded with close- and far-range microphones, with varying background noise levels and audio quality. The test set consists of 13078 utterances recorded by 142 different speakers, mapped to 70 unique intents. The audio recordings are labeled with their corresponding intent, given by the combination of action and scenario. The evaluation metric for the IC task is intent accuracy.

INTERACTIVE EMOTIONAL DYADIC MOTION CAPTURE (IEMOCAP) [22] is a widely used benchmark dataset for emotion recognition (ER) tasks in human-computer interaction research. The dataset consists of audiovisual recordings of naturalistic interactions between two actors engaged in scripted scenarios, resulting in over 12 hours of data. Ten actors were instructed to portray a range of emotional states, resulting in a diverse set of emotions. The dataset is labeled with discrete emotion labels (i.e., happiness, anger, sadness, frustration, and neutral state) and continuous arousal annotations (i.e., activation, valence, and dominance). These two labels offer complementary insights into the emotional expressions identified in the corpus. The public dataset is divided into five sessions (i.e., splits) that are generally evaluated separately using a 5-fold cross-validation approach. However, for the current study, we consider the compound of the test sets more appropriate, both to facilitate a more comprehensive model evaluation of the models and to augment the size of the evaluation set. Following standard procedure [29], we excluded the imbalanced emotion categories to ensure that the remaining four classes (neutral, happy, sad, angry) have a similar number of data points. As a result, our dataset consists of 4990 samples. The metric for the ER task is accuracy, which is commonly adopted for benchmarking ER tasks and the IEMOCAP dataset. Accuracy could not be the best evaluation option for imbalanced datasets. Other options, such as the Unweighted Average Recall (UAR), can be explored. Still, as our approach is metric agnostic, we can apply it to explore subgroups' performance for a generic performance measure, and we would obtain close insights when switching the metric.

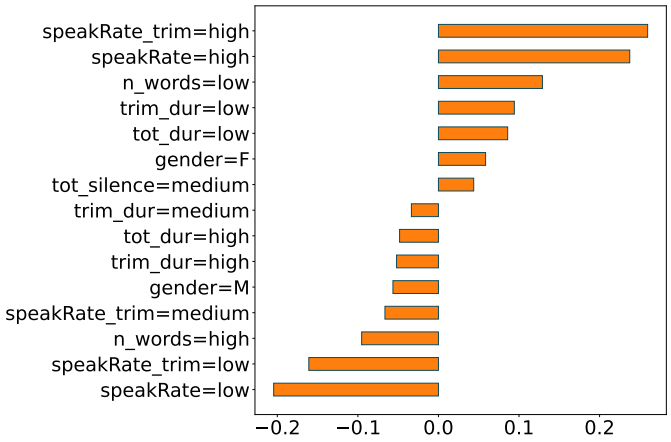


Fig. 10: RQ1. Top-15 Global Shapley values of WER divergence; LIBRISPEECH dataset, wav2vec 2.0 base model. Terms with positive contributions are associated with a WER higher than the WER on the entire dataset.

B. RQ1. How can we automatically identify and describe the most problematic subgroups for a given combination of SLU model, dataset, and task?

We summarize the impact of each item on the WER divergence using the global Shapley values, reported in Figure 10. Terms with positive contributions are associated with a WER higher than the WER on the entire dataset. Negative terms are associated with lower WER than average. Utterances with low speaking rates and many pauses are associated with a WER lower than the average. In contrast, low numbers of words and high speaking rates are associated with increased WER. Gender is an essential factor affecting model performance, with males having a lower WER than females. These findings underscore the significance of considering the speaking rate, pauses, and gender when designing and evaluating speech recognition models.

Figures 11(a) and 11(b) show the global Shapley Value of accuracy divergence for SLURP and IEMOCAP respectively. The action equal to quirk is the term that mostly globally affects SLURP performance. The 'field' also highly impacts the results. Utterances in the far-field are associated with lower than average performance, while close-field utterances with higher ones.

For IEMOCAP, we observe a high influence of the dataset-dependent metadata. Utterances with high valence, happiness as an emotion label, and low dominance are associated with lower performance than the average. In contrast, their oppo-

Dataset	Model	Negative Δ		Positive Δ	
		Baseline	Our	Baseline	Our
FSC	w2v2-b	-7.88	**-31.22**	6.32	**8.28**
	w2v2-l	-7.5	**-18.38**	**6.83**	**6.83**
	hub-b	-0.98	**-9.07**	**1.58**	**1.58**
	hub-l	-2.04	**-11.97**	**1.50**	**1.50**
SLURP	w2v2-b	**-19.50**	**-19.50**	7.56	**8.26**
	w2v2-l	**-17.41**	**-17.41**	8.11	**8.85**
	hub-b	-20.49	**-21.20**	7.61	**7.98**
	hub-l	-11.54	**-11.98**	6.27	**7.36**
IEMO	w2v2-b	-10.62	**-29.92**	10.95	**23.86**
	w2v2-l	-9.16	**-29.74**	6.03	**23.67**
	hub-b	-13.83	**-42.18**	9.62	**31.09**
	hub-l	-9.99	**-32.36**	10.62	**22.79**
LIBRISPEECH	w2v2-b	3.04	**11.24**	-1.37	**-2.79**
	w2v2-l	2.39	**8.74**	-0.64	**-2.05**
	hub-b	3.09	**9.90**	-0.98	**-2.83**
	hub-l	2.50	**7.30**	-0.65	**-1.68**

TABLE IX: RQ1. Maximum negative and positive divergence (Δ) for the baseline and our approach. Best results are highlighted in bold. Our approach is always superior or on par with the baseline. Note that for FSC, the maximum positive divergence is always similar, if not identical, since both approaches retrieve the subgroup(s) for which the model achieves 100% accuracy.

sites (high valence, sad as emotion label, and high dominance) are associated with higher performance.

Comparison with baselines Table IX compares our approach with one-level subgroup identification. Our approach consistently demonstrates superior or comparable performance when compared to the baseline. This highlights the effectiveness and strength of our approach in addressing the research problem at hand and further support the validity of our proposed method in subgroup identification tasks.

C. RQ2. What is the effect of the model size on subgroup performance? Does The large the better hold true?

Table X provides detailed information about the subgroups that exhibit the most significant performance improvements and decreases when scaling up the HuBERT model size. These subgroups represent specific characteristics or conditions within the datasets where scaling up the model has a notable impact.

Figure 12 presents the distribution of the cross-model performance gap for all the considered datasets. This figure visualizes the performance gap between different size versions of HuBERT, showcasing the differences in performance achieved by scaling it up. By examining the distribution, one can gain insights into the overall impact and effectiveness of scaling up the model across the datasets.

FSC. Both HuBERT base and large exhibit similar performance on this dataset, achieving accuracies of 98.42% and 98.50%, respectively. However, when scaling up the model, we observed an improvement in performance for 51.33% of the examined subgroups, while 32.97% experienced a decrease. The initial section of Table X highlights the subgroups with the most significant increase (by 9.84%) and decrease (by -10.64%) in performance.

SLURP. Regarding HuBERT, for 81.78% of the explored subgroup, performance increases from base to large. Their overall accuracy is 87.70% and 89.25%, respectively. The subgroups with the most significant increase in performance (12.70%) and the largest decrease (-3.60%) are shown in the second block of Table X.

IEMOCAP. When scaling HuBERT, overall performance rises from 67.44% to 74.99%. The improvement is also at the subgroup level. For almost all the explored subgroups (93.95%), performance improves from base to large, confirming the expected behavior when scaling up the size. The highest increase is by 27.92%. Still, for some subgroups, we observe a performance decrease. The highest decrease is by -6.43%, where accuracy drops from 76.43% to 70.00%.

LIBRISPEECH. The HuBERT base version achieves an overall WER score of 6.56%, while the large a much lower (thus, better) 3.50%. Most importantly, HuBERT large behaves *better* than base on 100% of the explored subgroups. Hence, HuBERT large shows *better performance both overall and at the subgroup level*. The highest improvement is -6.16% for the subgroup {*"gender=female, speaking rate=high, trimmed speaking rate=high, trimmed duration=low"*}. While we observe a significant improvement for this subgroup, the large model still underperforms overall performance, revealing that this subgroup is still more difficultly modeled.

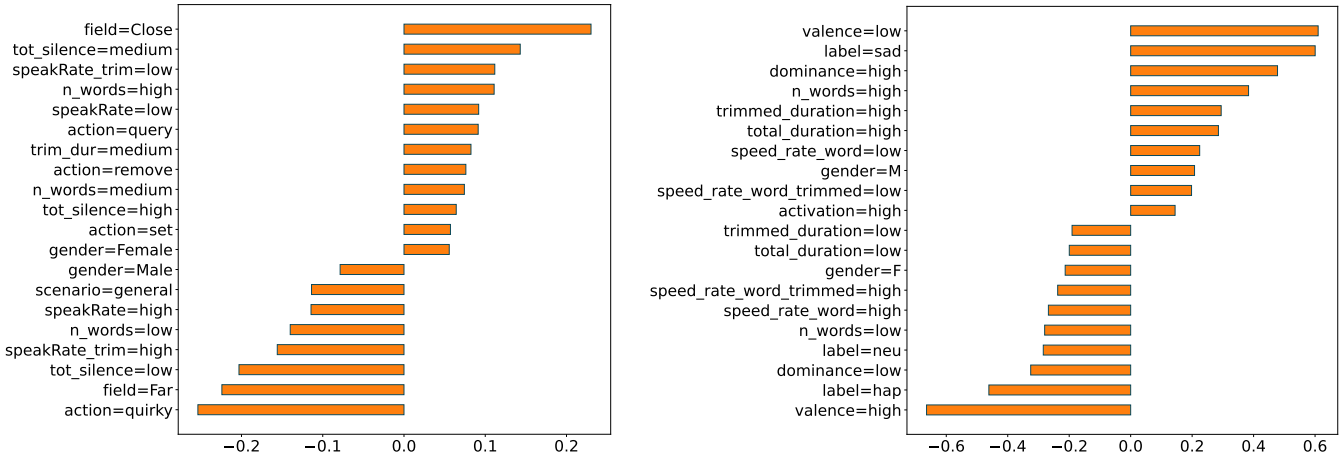
Summary of findings. Our findings demonstrate that scaling up the HuBERT model yields benefits for the majority of the analyzed subgroups across different datasets. However, the extent of improvement varies. In some cases, such as LIBRISPEECH, the improvement is observed for all explored subgroups, while in others, such as IEMOCAP and SLURP, it is significant for a high percentage of subgroups (93.95% and 81.78%, respectively). Conversely, in the FSC dataset, the improvement is less pronounced, impacting only 51.33% of the subgroups.

D. RQ3. Is the performance bias on specific subgroups independent of the model architecture?

Table XI outlines the subgroups with the highest performance improvement and the highest decrease when changing the models' architecture from wav2vec 2.0 to HuBERT large.

Figure 13 provides a visual representation of the distribution of the cross-model performance gap when changing the architecture from wav2vec 2.0 to HuBERT large. It illustrates the differences in performance achieved by transitioning from one model to another for each of the explored datasets.

FSC. In general, HuBERT large demonstrates superior performance compared to wav2vec 2.0 large, 98.50% vs. 93.17%, respectively. When considering specific subgroups, transitioning from wav2vec 2.0 to HuBERT leads to performance improvements in 91.84% of the subgroups, while only 5.18% experience a decrease. The largest increase (by 24.43%) and decrease (by -7.80%) in performance are documented in the initial block of Table XI. For the FSC dataset, gender is the

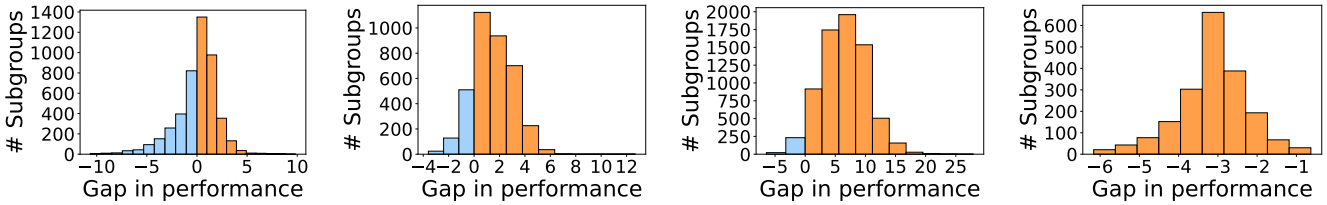


(a) SLURP dataset. (b) IEMOCAP dataset.

Fig. 11: RQ2. Gap contribution when scaling up HuBERT, considering SLURP and IEMOCAP datasets.

Dataset		*Subgroups*	*Sup*	gap_f	$f_{\texttt{hub-b}}$	$f_{\texttt{hub-l}}$	*t*
FSC	↑	{*"age=22-40, gender=male, num words=medium, tot silence=high"*}	0.03	9.84	89.34	99.18	3.17
	↓	{*speaking rate=low, trimmed speaking rate=low, tot silence=low, trimmed duration=low*}	0.04	-10.64	97.16	86.52	3.21
SLURP	↑	{*"field=far, scenario=general"*}	0.03	12.69	66.50	79.19	4.04
	↓	{*"scenario=qa, duration=high"*}	0.03	-3.60	89.45	85.85	1.57
IEMO	↑	{*"label=sad, activation=high"*}	0.03	27.92	51.30	79.22	5.35
	↓	{*gender=female, activation=medium, trimmed speaking rate=high, label=neutral*}	0.03	-6.43	76.43	70.00	1.21
LS	↑	{*gender=female, speaking rate=high, trimmed speaking rate=high, trimmed duration=low*}	0.04	-6.16	16.26	10.10	2.04

TABLE X: RQ2. Gap for performance measure f when scaling up the HuBERT size from base (90 million parameters) to large (300 million parameters). (↑) denotes the highest performance improvement, (↓) indicates the largest decrease. The *t* column indicates the Welch's t-test value.



(a) FSC. Performance increases for 51.33% of the explored subgroups and decreases for 32.97% of them.

(b) SLURP. Performance increases for 81.78% of the explored subgroups and decreases for 15.79% of them.

(c) IEMOCAP. Performance increases for 95.89% of the explored subgroups and decreases for 2.52% of them.

(d) LIBRISPEECH. Performance increases for 100% of the explored subgroups.

Fig. 12: RQ2. Gap contribution when scaling up HuBERT, considering different datasets.

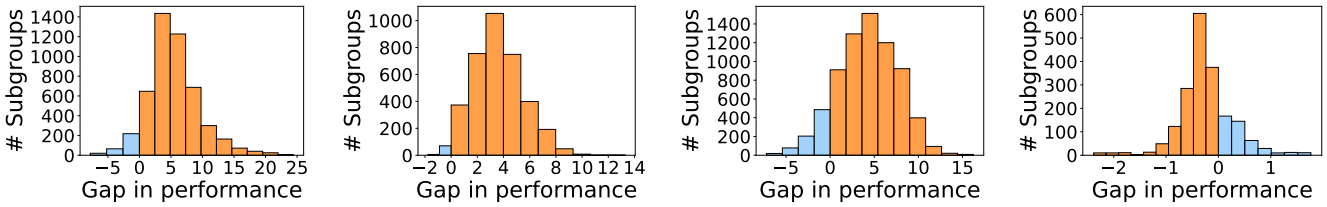
most influential factor in the global Shapley values of accuracy gap from wav2vec 2.0 to HuBERT large (Fig. 14), with males benefiting more than females. An interesting observation can be made regarding the influence of silence on performance. Utterances with a high number of silences tend to result in superior performance compared to the average, while utterances with a low number of silences tend to yield lower performance. Additionally, the action "activate" consistently exhibits a positive influence, with higher accuracy than the average, while the action "deactivate" shows the opposite pattern, indicating lower performance.

SLURP. Changing from wav2vec 2.0 to HuBERT proves to be advantageous at the overall level (85.59% vs. 89.25%) but also for the majority of the analyzed subgroups (97.43%). Detailed information can be found in the second portion of Table XI, where the most significant improvement reaches 13.30%, while the largest decrease is a mere −1.78% for a subgroup in which both models still perform above the average performance.

IEMOCAP. When transitioning from wav2vec 2.0 to Hu-

Dataset		*Subgroups*	*Sup*	gap_f	$f_{\texttt{w2v2-l}}$	$f_{\texttt{hub-l}}$	t
FSC	↑	{*"action=increase, gender=male, speaking rate=high"*}	0.03	24.43	74.81	99.24	6.15
	↓	{*"speaking rate=low, trimmed speaking rate=low, tot silence=low, trimmed duration=low"*}	0.03	-7.80	94.33	86.52	2.18
SLURP	↑	{*"action=remove, num words=low"*}	0.03	13.30	82.90	96.20	6.40
	↓	{*"action=query, language=other, speaking rate=medium, trimmed speaking rate=medium"*}	0.03	-1.78	92.13	90.35	0.87
IEMO	↑	{*"label=sad, activation=high"*}	0.03	16.23	62.99	79.22	3.17
	↓	{*"gender=male, label=neutral, speaking rate=medium, valence=medium"*}	0.04	-7.22	78.89	71.67	3.77
LS	↑	{*"speaking rate=high, trimmed speaking rate=high, tot duration=low, tot silence=low, trimmed duration=low"*}	0.05	-2.38	12.53	10.14	0.78
	↓	{*"trimmed speaking rate=high, tot duration=low, tot silence=medium, trimmed duration=low"*}	0.03	1.77	7.58	9.35	0.64

TABLE XI: RQ3. Gap for performance measure f when changing the models' architecture from wav2vec 2.0 to HuBERT large. (↑) denotes the highest performance improvement, (↓) indicates the largest decrease. The t column indicates the Welch's t-test value.



(a) FSC. Performance increases for 91.84% of the explored subgroups and decreases for 5.18% of them.

(b) SLURP. Performance increases for 97.43% of the explored subgroups and decreases for 1.51% of them.

(c) IEMOCAP. Performance increases for 89.13% of the explored subgroups and decreases for 8.60% of them.

(d) LIBRISPEECH. Performance increases for 79.20% of the explored subgroups and decreases for 20.59% of them.

Fig. 13: RQ3. Gap distribution when changing wav2vec 2.0 to HuBERT large.

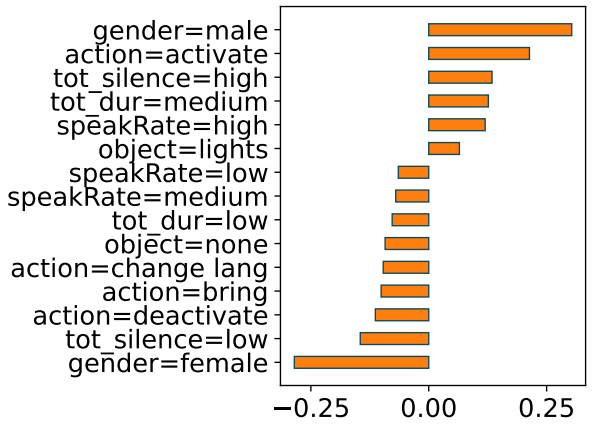


Fig. 14: Global Shapley values of accuracy gap. FSC dataset, wav2vec 2.0 to HuBERT large model.

BERT large, performance improves for 89.13% of the examined subgroups, while 8.60% experience a decrease. We recall that wav2vec 2.0 large attained an overall accuracy of 71.18%, whereas HuBERT achieved a higher accuracy of 74.99%. Referencing the third section of Table XI, we observe the largest increase in performance (by 16.23%) and the largest decrease (by −7.22%).

LIBRISPEECH. On this dataset, wav2vec 2.0 and HuBERT exhibit similar performance, with WER of 3.82% and 3.50%, respectively. However, when shifting from wav2vec 2.0 to HuBERT large, we observe performance improvements in 79.20% of the analyzed subgroups, while 20.59% undergo a decrease. Notably, the maximum increase (−2.88%) and decrease (1.77%) in performance, highlighted in the final section of Table XI, are relatively comparable.

Summary of findings. In contrast to our findings when transitioning from wav2vec base to HuBERT base, where different datasets exhibited varying effects on subgroups (with positive benefits for FSC and SLURP, but negative impacts on most subgroups in IEMOCAP and LIBRISPEECH), the transition from one large version to the other generally leads to performance improvements for most subgroups across all considered datasets. The percentage of subgroups benefiting from the architecture change ranges from a minimum of 79.20% for LIBRISPEECH to a maximum of 97.43% for SLURP. These results indicate a more consistent positive impact on performance when upgrading from one large model version to another across the analyzed datasets.

E. RQ4. Are multilingual SLU models more sensitive to subgroup performance bias than monolingual ones?

Figure 15 reports the top-15 items with the highest Global Shapley value (in absolute terms) of the accuracy gap from wav2vec 2.0 large mono-lingual to XLSR-53. The action "activate" has the most significant global impact on the performance of FSC. Additionally, the number of words in the utterances has a substantial influence on the results. Utterances with a higher word count tend to yield lower-than-average performance, while those with a moderate number

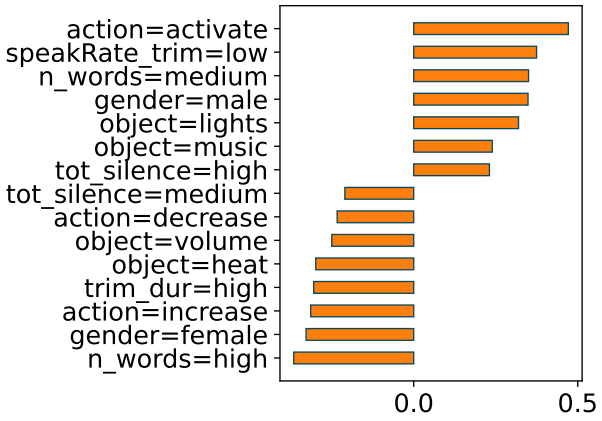


Fig. 15: RQ4. Global Shapley values of accuracy gap for wav2vec 2.0 large mono-lingual to XLSR-53, top-15. FSC.

of words result in higher performance. Moreover, gender plays a prominent role, with utterances from female speakers being associated with lower performance compared to the average, while utterances from male speakers exhibit higher performance.