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.

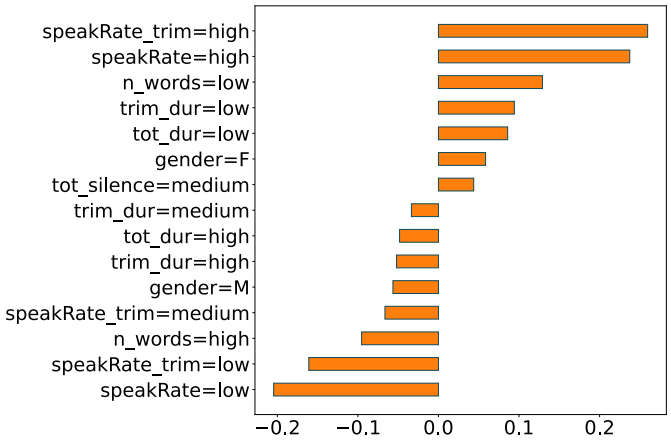


Fig. 10: RQ1. Top-15 Global Shapley values of WER divergence; LIBRISPEECH dataset, wav2vec 2.0 base model. Terms with positive contributions are associated with a WER higher than the WER on the entire dataset.

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 Baseline	Negative Δ Our	Positive Δ Baseline	Positive Δ 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.

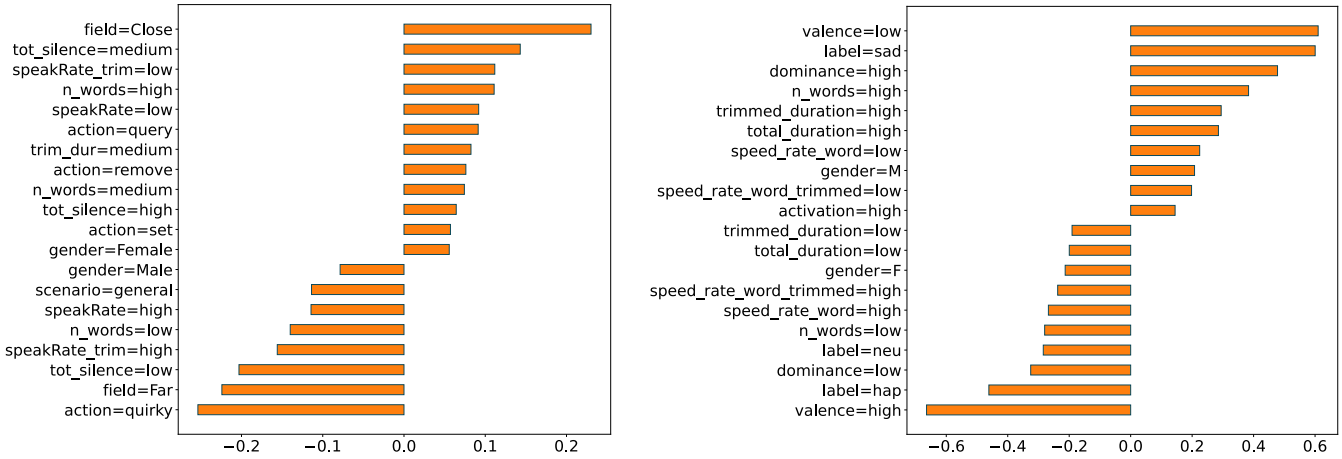
Summary of findings. Our findings demonstrate that scaling up the HuBERT model yields benefits for the majority of the analyzed subgroups across different datasets. However, the extent of improvement varies. In some cases, such as LIBRISPEECH, the improvement is observed for all explored subgroups, while 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

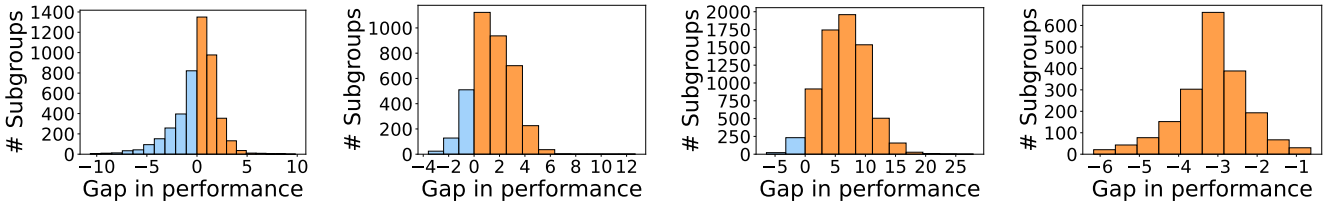


(a) SLURP dataset. (b) IEMOCAP dataset.

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.

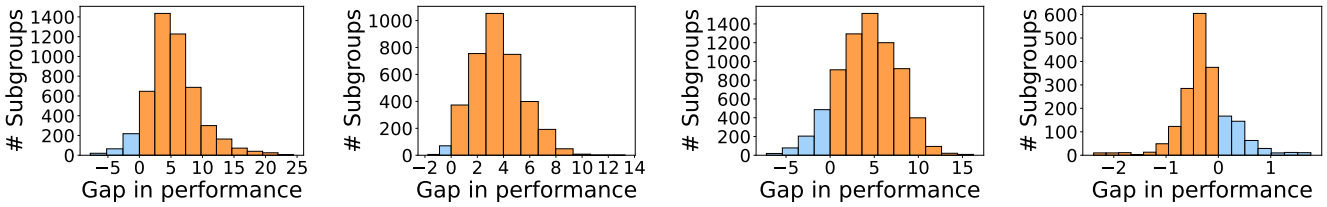
most influential factor in the global Shapley values of accuracy gap from wav2vec 2.0 to HuBERT large (Fig. 14), with males benefiting more than females. An interesting observation can be made regarding the influence of silence on performance. Utterances with a high number of silences tend to result 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.

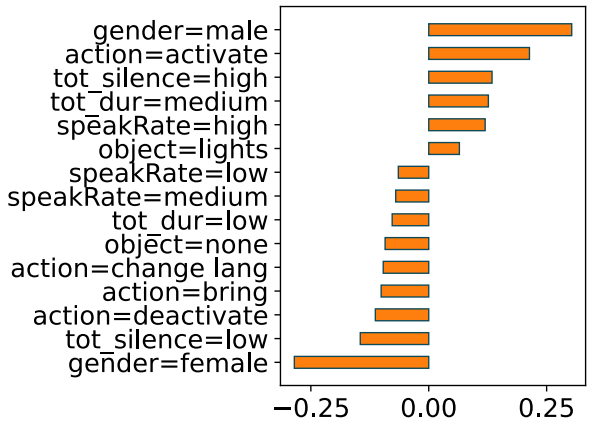


Fig. 14: Global Shapley values of accuracy gap. FSC dataset, wav2vec 2.0 to HuBERT large model.

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

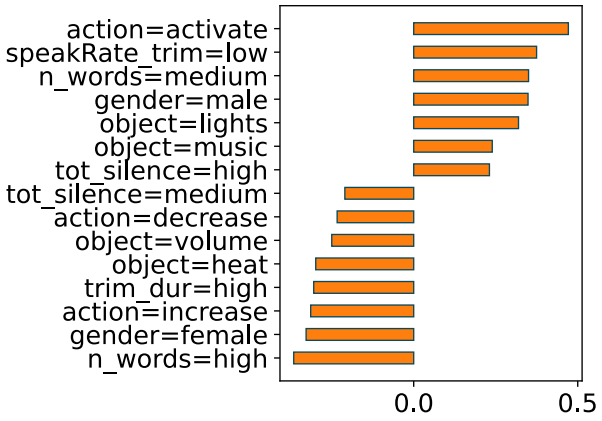


Fig. 15: RQ4. Global Shapley values of accuracy gap for wav2vec 2.0 large mono-lingual to XLSR-53, top-15. FSC.

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.