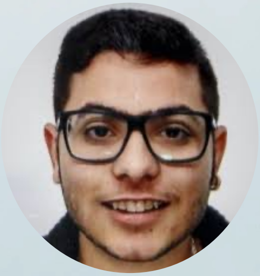


Assessing Speech Model Performance: A Subgroup Perspective

Alkis Koudounas, **Eliana Pastor**, Elena Baralis

Politecnico di Torino - Italy

SEBD 2024

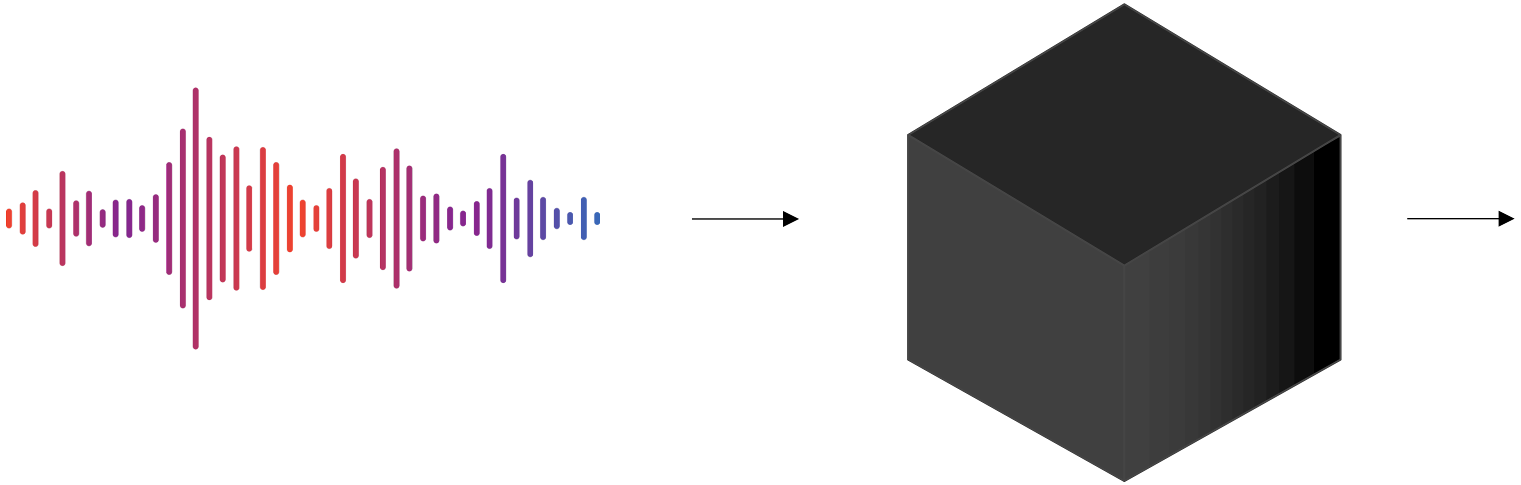
Politecnico di Torino 1859

DBMG

ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

In collaboration with amazon | science

Our scenarios



Automatic Speech Recognition

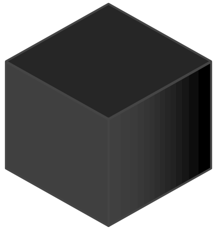
Turn on the kitchen lights

Intent classification

Action: activate
Object: lights
Location: kitchen

Emotion recognition

Neutral



PERFORMANCE
X%

LOW ERROR RATE

HIGH ERROR RATE

AS OVERALL
MODEL
BEHAVIOR

Outline

- **Identification of interpretable subgroups** with divergences in performance

- **Model comparison** from the subgroup perspective

- **Subgroup-guided acquisition** for model improvement

Outline

- **Identification of interpretable subgroups** with divergences in performance

- **Model comparison** from the subgroup perspective

- **Subgroup-guided acquisition** for model improvement

How to make an interpretable data grouping?

Enhance utterance with interpretable metadata

Speaker demographics

Speaking and recording conditions

Task- or dataset specific features

Metadata

gender=female
country=Italian
noise-level=high
speaking rate=fast
…

Subgroup identification

- **Automatic** identification of subgroups via **frequent pattern mining**
  - Slicing in the interpretable attibute space

- Compute subgroup **divergence**

performance measure

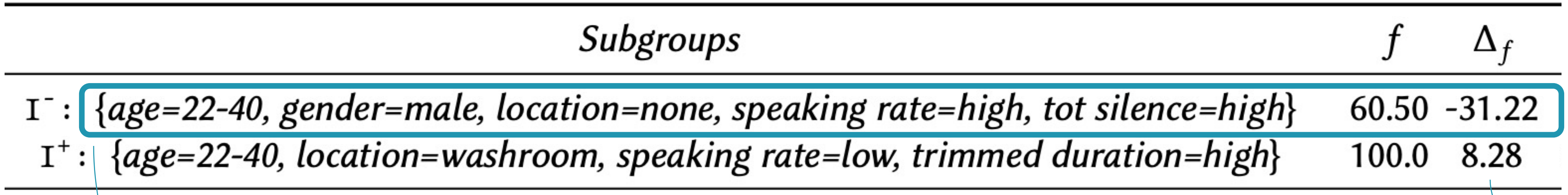
$$\Delta(S) = f(S) - f(D)$$

pattern, e.g., {age=20-35, gender=female}

all dataset

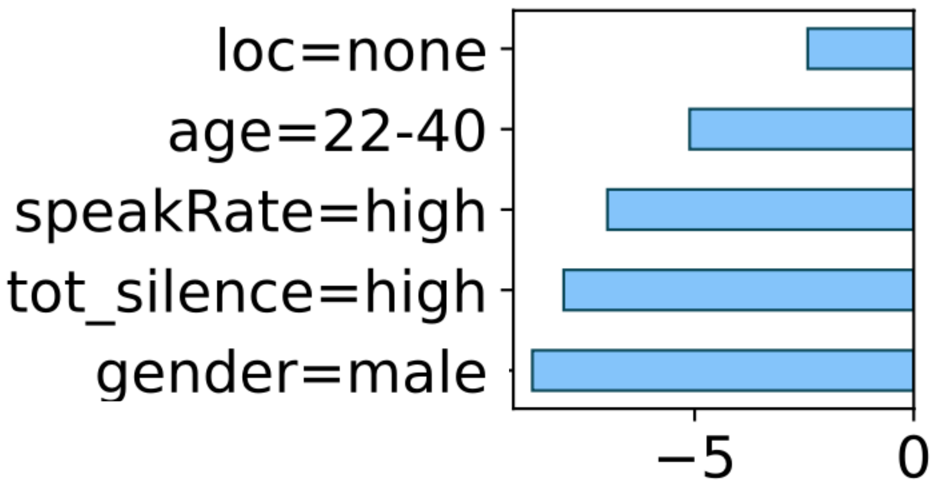
Divergent subgroup

By 31.22 less accurate!



	Subgroups	f	Δ_f
$\mathtt{I}^-$:	*{age=22-40, gender=male, location=none, speaking rate=high, tot silence=high}*	60.50	-31.22
$\mathtt{I}^+$:	*{age=22-40, location=washroom, speaking rate=low, trimmed duration=high}*	100.0	8.28

More accurate than average

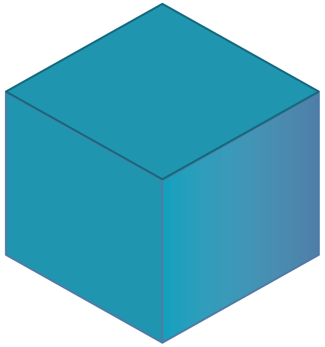


Outline

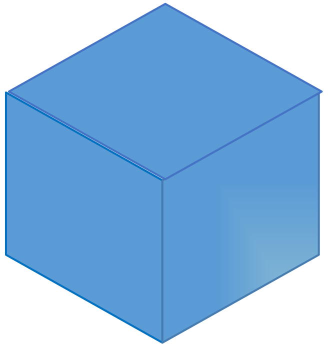
- **Identification of interpretable subgroups** with divergences in performance

- **Model comparison** from the subgroup perspective

- **Subgroup-guided acquisition** for model improvement



Accuracy 91.72%



Accuracy 93.17%

Which model to choose?

.. most accurate..?

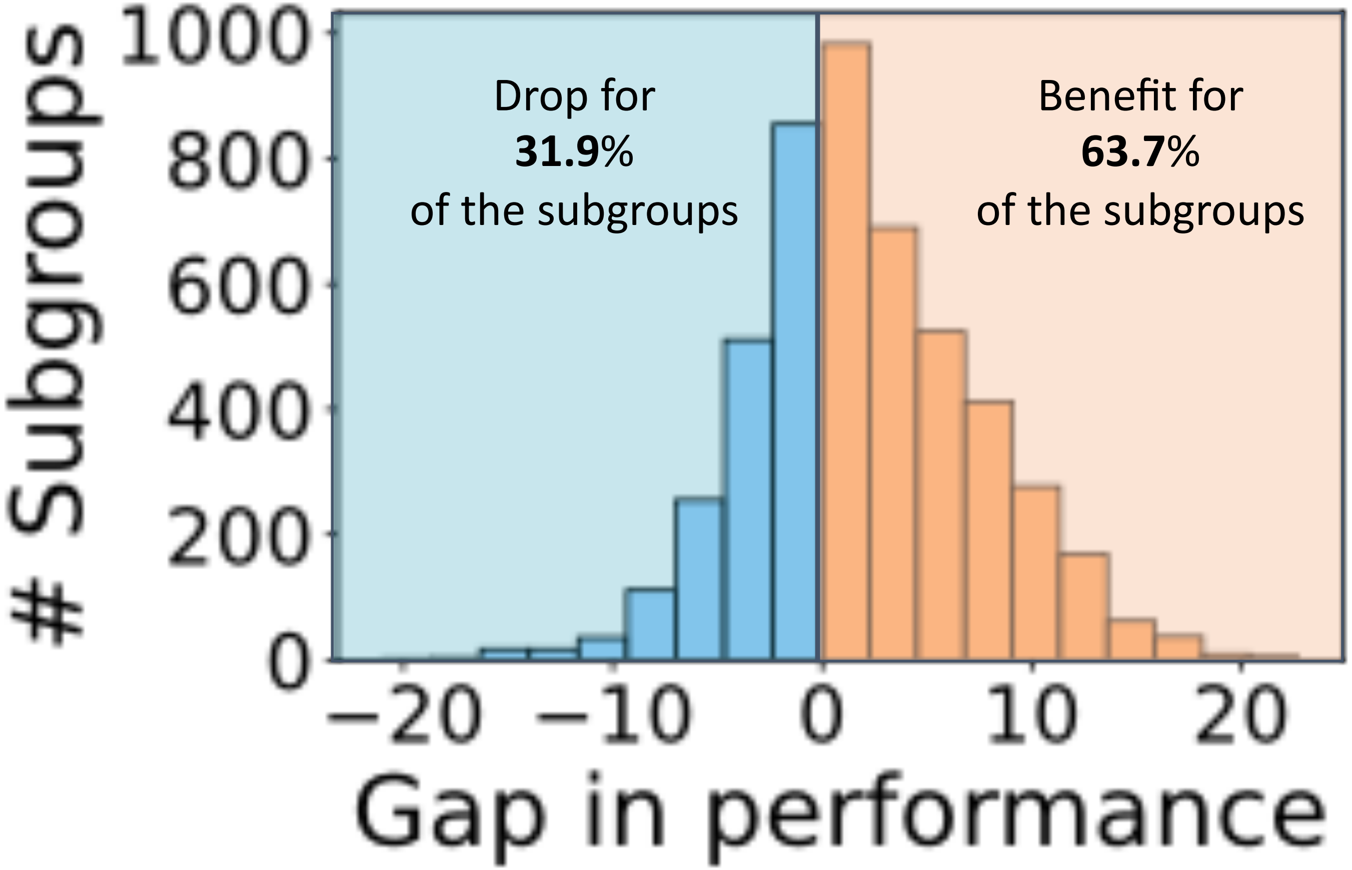
But on subgroups?

Inter-model performance gap

S = **pattern**, e.g., {age=20-35, gender=female}

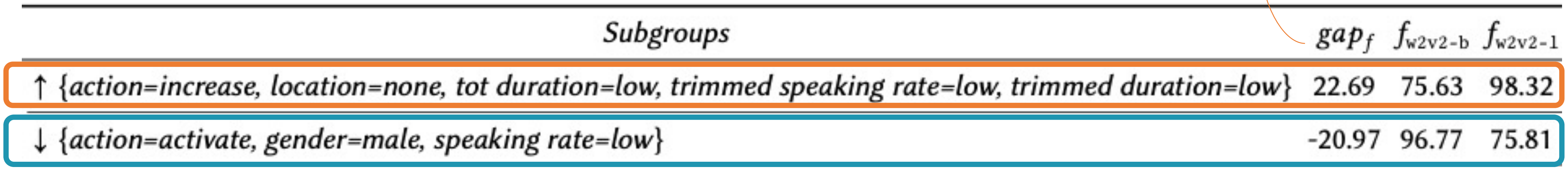
$$gap_f(S, M_1, M_2) = \underbrace{f(S, M_2)}_{\textbf{Performance on S} \text{ of model } \boldsymbol{M_2}} - \underbrace{f(S, M_1)}_{\textbf{Performance on S} \text{ of model } \boldsymbol{M_1}}$$

Distribution of gain in performance



An example

Increase in performance



Subgroups	gap_f	$f_{\text{w2v2-b}}$	$f_{\text{w2v2-l}}$
↑ {*action=increase, location=none, tot duration=low, trimmed speaking rate=low, trimmed duration=low*}	22.69	75.63	98.32
↓ {*action=activate, gender=male, speaking rate=low*}	-20.97	96.77	75.81

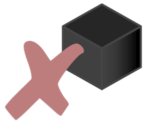
Drop in performance

Outline

- **Identification of interpretable subgroups** with divergences in performance
- **Model comparison** from the subgroup perspective
- **Subgroup-guided acquisition** for model improvement

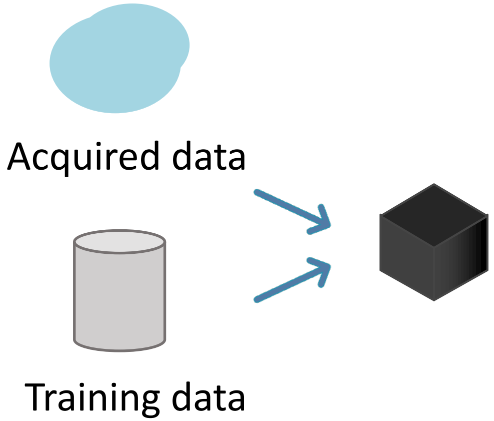
Subgroup-guided data acquisition

Speaking rate=high, gender=male



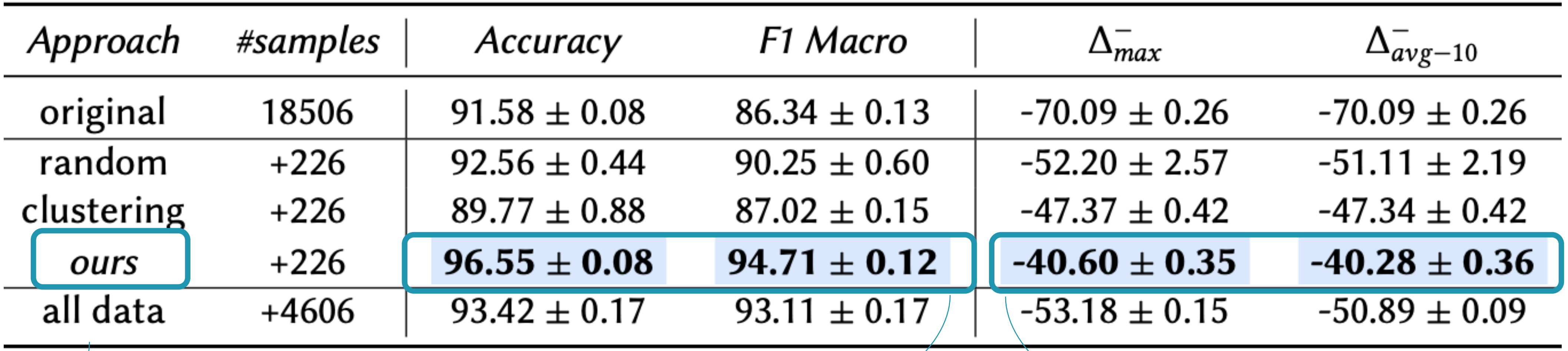
Step 1.
Identify the divergent patterns

Step 2.
Acquire data satisfying the patterns



Step 3.
Speech model re-training

Results of subgroup-guided data acquisition



Approach	*#samples*	*Accuracy*	*F1 Macro*	Δ^{-}_{max}	Δ^{-}_{avg-10}
original	18506	91.58 ± 0.08	86.34 ± 0.13	-70.09 ± 0.26	-70.09 ± 0.26
random	+226	92.56 ± 0.44	90.25 ± 0.60	-52.20 ± 2.57	-51.11 ± 2.19
clustering	+226	89.77 ± 0.88	87.02 ± 0.15	-47.37 ± 0.42	-47.34 ± 0.42
ours	+226	**96.55 ± 0.08**	**94.71 ± 0.12**	**-40.60 ± 0.35**	**-40.28 ± 0.36**
all data	+4606	93.42 ± 0.17	93.11 ± 0.17	-53.18 ± 0.15	-50.89 ± 0.09

Improvement compared to acquire all the data!

We improve overall performance!

We improve subgroup performance!

✉ eliana.pastor@polito.it

eliana__pastor

elianap.github.io/

or let’s have a chat! ☀