Dataset Curation for Visual Speech Recognition

Claire Chen, Maya Krolik

The problem: lack of high-quality lip-reading datasets and metric

- The process of collecting and processing training data is tedious

Source: Pingchuan Ma et al. *Auto-avsr: Audio-visual speech recognition with automatic labels. arXiv:2303.14307, 2023*

The problem: lack of high-quality lip-reading datasets and metric

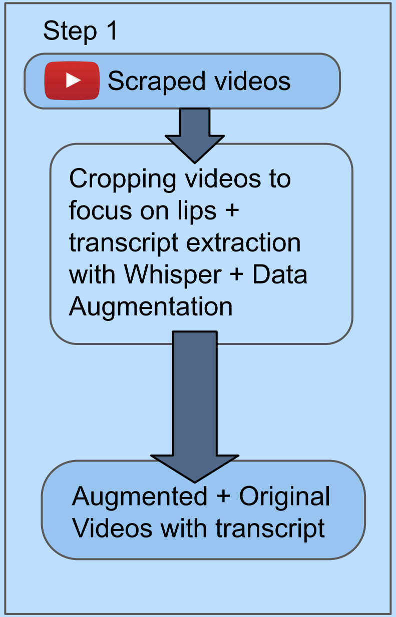
- The process of collecting and processing training data is tedious
- Lack of standardized quality control metrics in datasets
    - Dataset lacks diversity and trained models do not generalize to the real world

Source: Yao et al. *A comprehensive multimodal dataset for contactless lip reading and acoustic analysis*

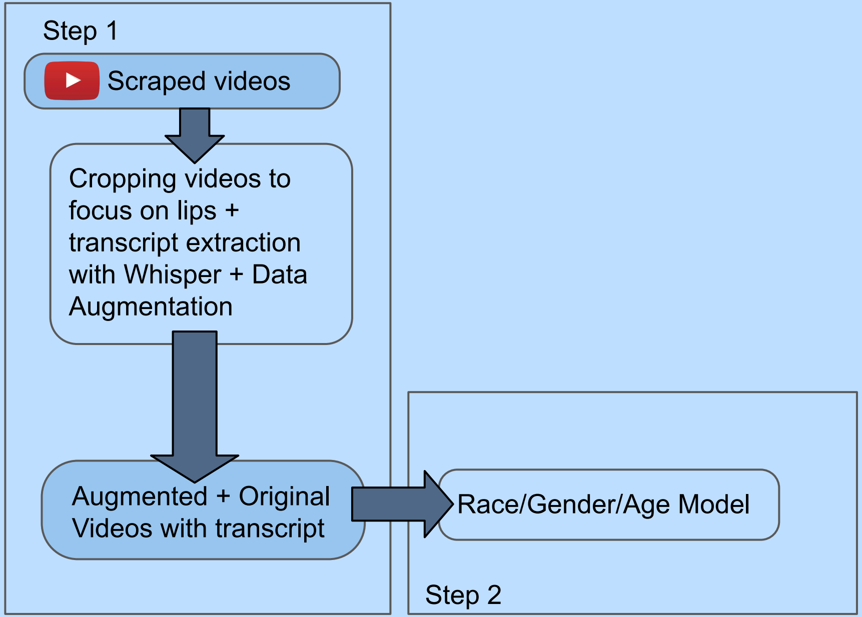
The solution: scalable, customizable pipeline for curating diverse datasets

- Develop more accurate models generalizable to the real world

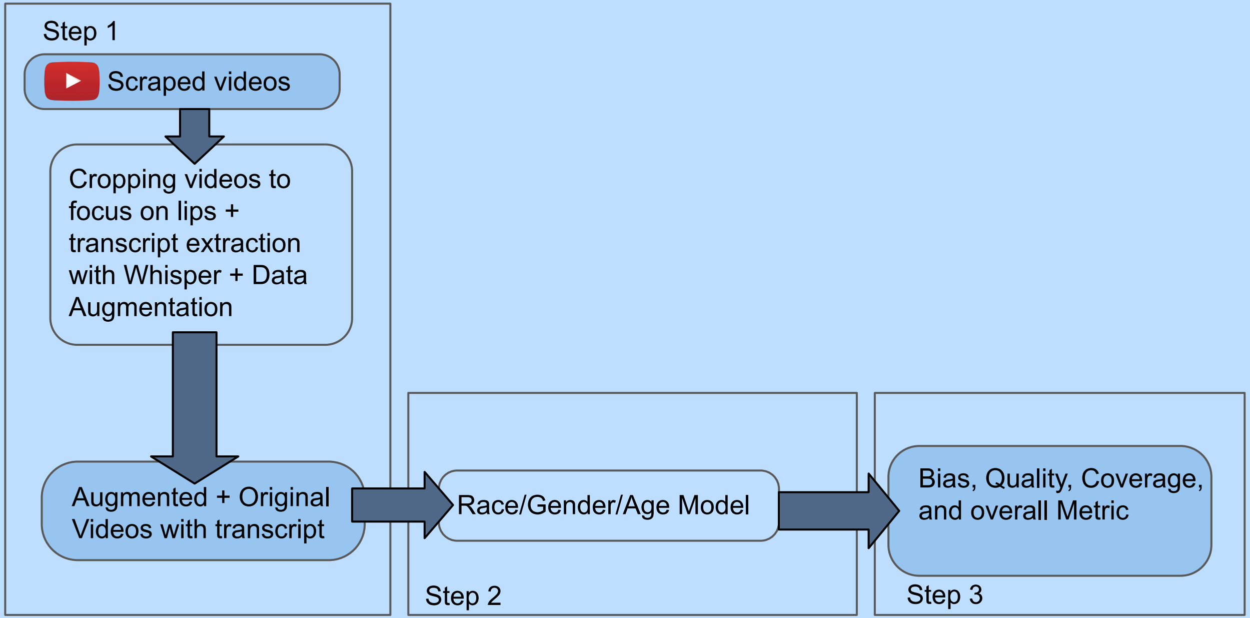
Dataset Curation Pipeline



Dataset Curation Pipeline



Dataset Curation Pipeline



Metric

- Goal: ensure balanced representation across demographic categories
    - Coverage score

Metric: cross-category coverage score

- Systematically evaluates all possible demographic intersections (i.e. every combination of race, gender, and age)
    - (White, Female, Child)
    - (Asian, Male, Senior)
    - …

Metric: cross-category coverage score

- Count the raw number of samples for each group
- Normalize each group’s count by dividing it by the largest count among all groups

Metric: cross-category coverage score

- Compute the raw number of samples for each group
- Normalize each group's count by dividing it by the largest count among all groups

Category	Raw count	Normalized coefficient
(White, Male)	1	
(White, Female)	2	
(Asian, Male)	1	
(Asian, Female)	0	

Metric: cross-category coverage score

- Compute the raw number of samples for each group
- Normalize each group's count by dividing it by the largest count among all groups

Category	Raw count	Normalized coefficient
(White, Male)	1	0.5
(White, Female)	2	1
(Asian, Male)	1	0.5
(Asian, Female)	0	0

Metric: cross-category coverage score

- Coverage score = 0.5 * minimum normalized coefficient + 0.5 * average normalized coefficient

Metric: cross-category coverage score

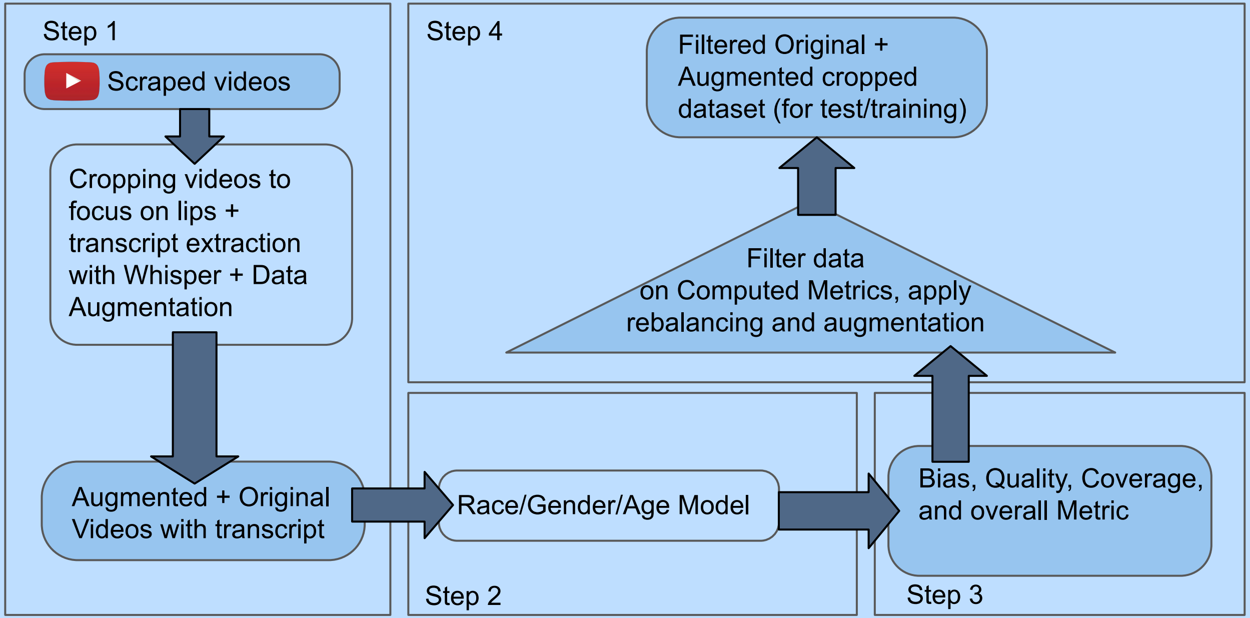
- Coverage score = 0.5 * minimum normalized coefficient + 0.5 * average normalized coefficient
- e.g. coverage score = 0.5 * 0 + 0.5 * 0.5 = 0.25

Category	Raw count	Normalized coefficient
(White, Male)	1	0.5
(White, Female)	2	1
(Asian, Male)	1	0.5
(Asian, Female)	0	0

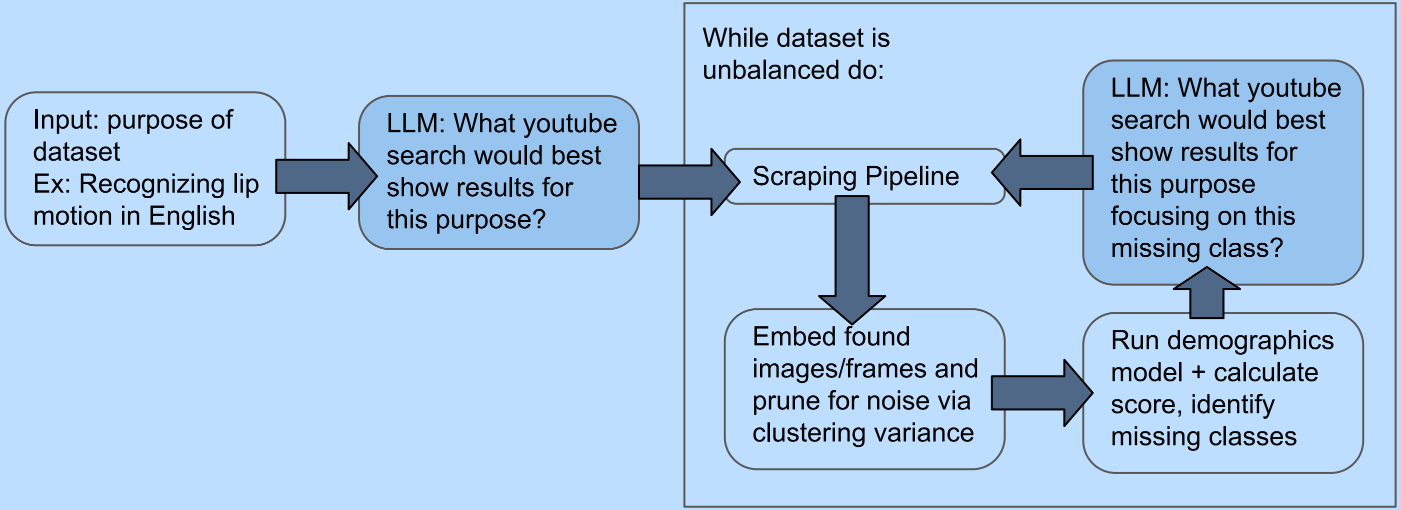
Metric: minimum representation constraint

- Minimum representation constraint: number of samples required in each category

Dataset Curation Pipeline

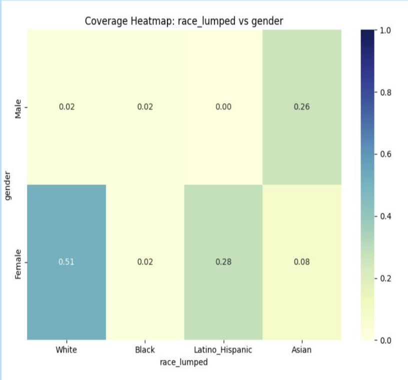


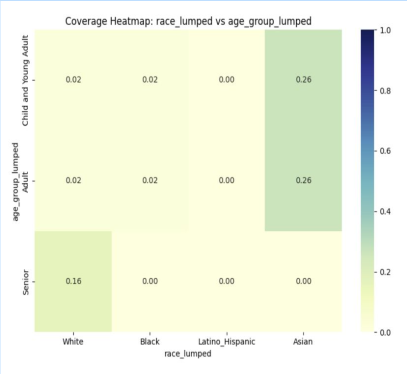
Dataset Curation Pipeline

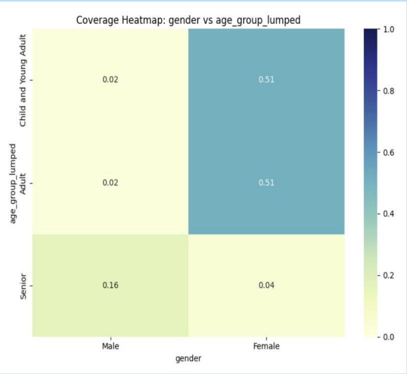


Results: coverage score pre-LLM = 0.07

Before pipeline

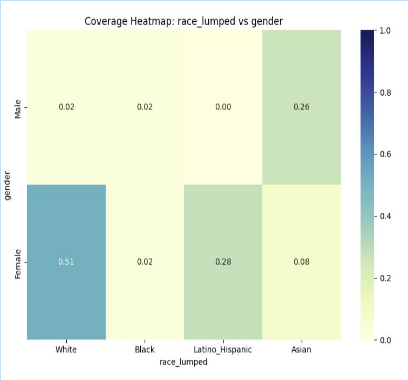


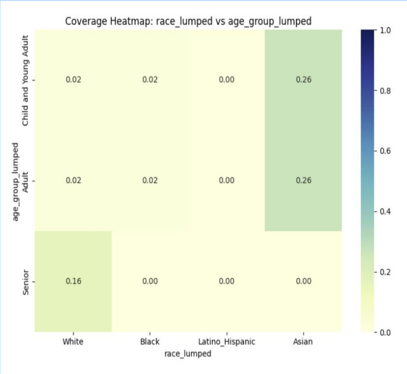


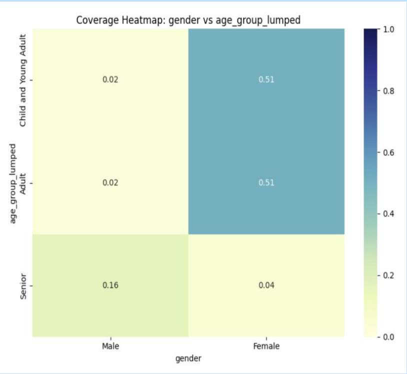


Results: coverage score post-LLM = 0.11

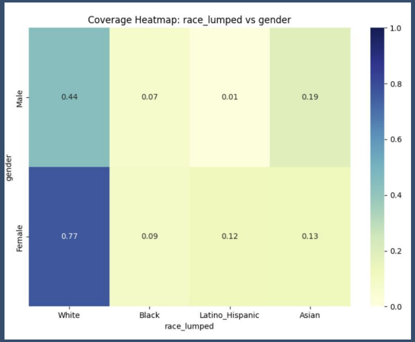
Before pipeline

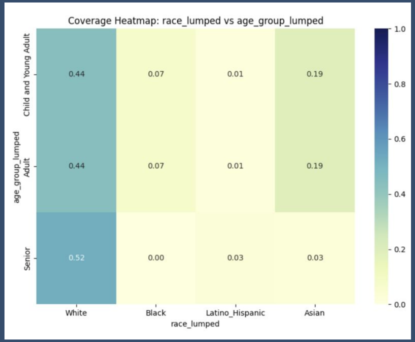


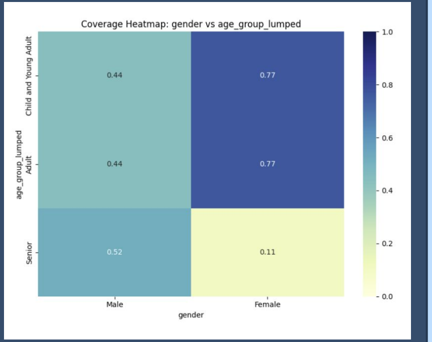




After pipeline

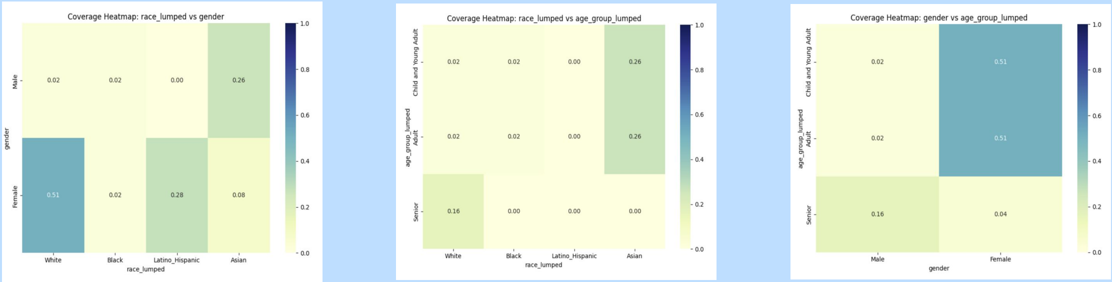




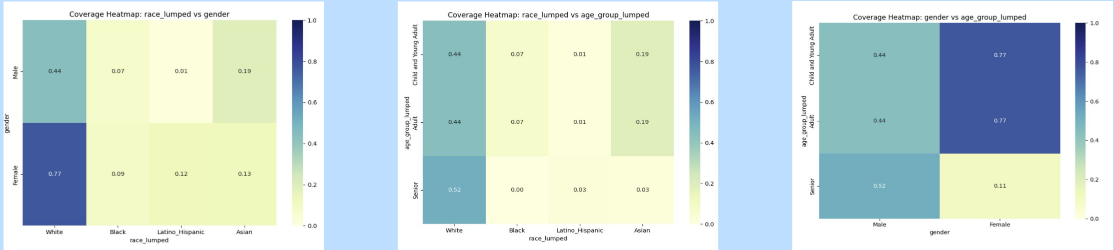


Results: baseline coverage score = 0.06

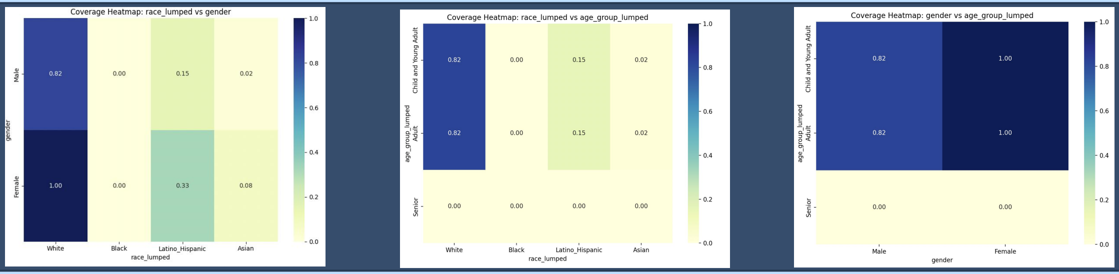
Before pipeline



After pipeline



Baseline Model



Results cont. and Insights

Results:

- Working pipeline and metric
- Proof of concept that LLMs iteratively improve dataset quality
- Applied metric to existing dataset to show lack of demographic coverage

Results cont. and Insights

Results:

- Working pipeline and metric
- Proof of concept that LLMs iteratively improve dataset quality
- Applied metric to existing dataset to show lack of demographic coverage

Given more time we would:

- Fine-tune our LLM to produce better prompts

Results cont. and Insights

Results:

- Working pipeline and metric
- Proof of concept that we can use LLMs to iteratively improve dataset quality
- Applied metric to existing dataset to show lack of demographic coverage

Given more time we would:

- Fine-tune our LLM to produce better prompts

Insights:

- Had a lot of fun learning how to combine multiple CV models to create a usable product
- Validating ideas on a smaller dataset before scaling up helps identify issues early

Thank you!

Claire Chen, Maya Krolik

Links

Paper submission: https://drive.google.com/file/d/1KtA4QkXR0Y4JqFiQvd7Mg48jgq374SMg/view?usp=sharing

Code: https://github.com/mayakrolik/6.S058-Final-Project

Dataset: https://www.kaggle.com/datasets/lamayonesa/vsr-automatic-dataset