Don't speak too fast: The impact of data bias on self-supervised speech models

Yen Meng, Yi-Hui Chou, Andy T. Liu , Hung-yi Lee

National Taiwan University

AAAI-22

Outline

- Background & Motivation
- Introduction
- Methodology
- Experiments
- Conclusions

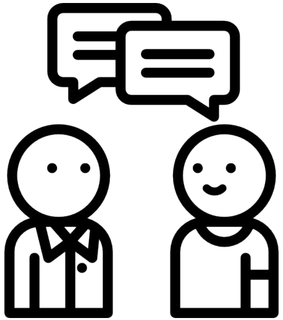
Outline

- Background & Motivation
- Introduction
- Methodology
- Experiments
- Conclusions

In real world applications, our collected audio data can be biased in different aspects:

Demographic: gender, age, accent, …

Content: topic, word use, …



Prosody: speech rate, tone, …

Data bias have become more aware in recent research

There are previous works investigating data bias of a single downstream task, such as ASR, speaker recognition, or speech translation

However, data bias in self-supervised pre-training is unexplored

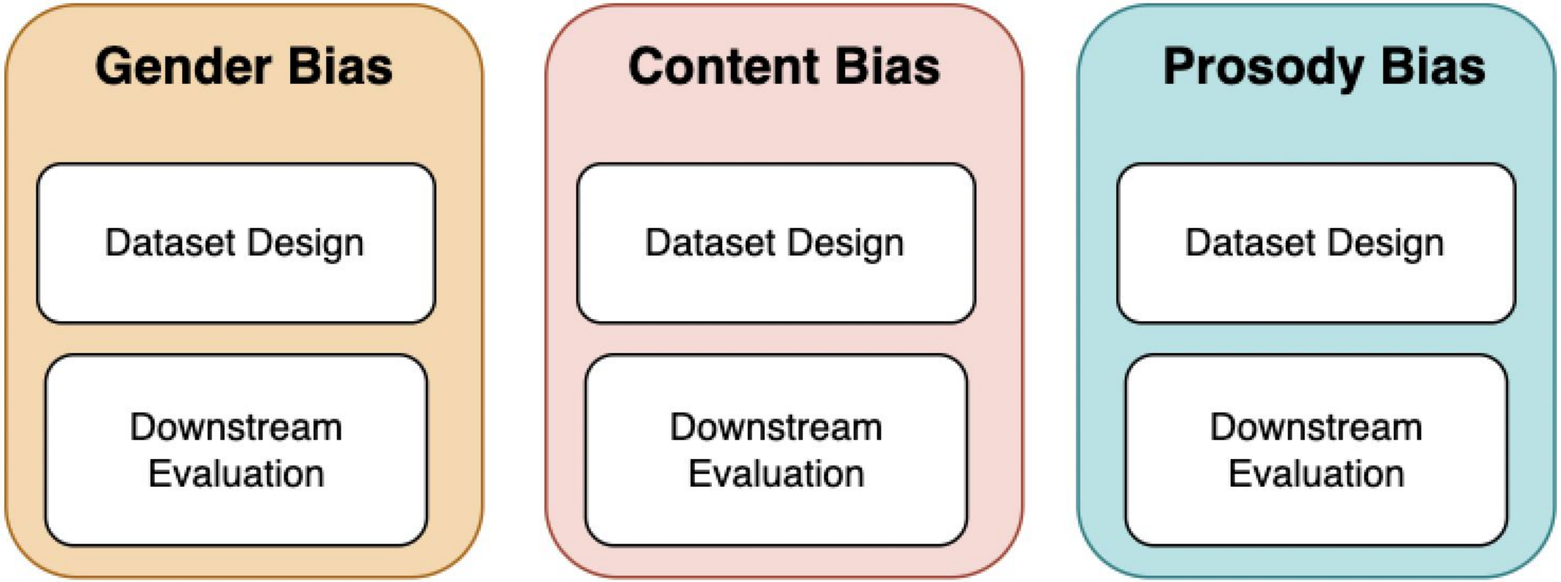
S3Ms are often pre-trained on “standard” datasets such as LibriSpeech

- How would data bias in pre-training affect S3Ms?
- Is “balanced” data for pre-training necessary to achieve generalizable performance in downstream tasks?

Outline

- Background & Motivation
- Introduction
- Methodology
- Experiments
- Conclusions

3 aspects of bias

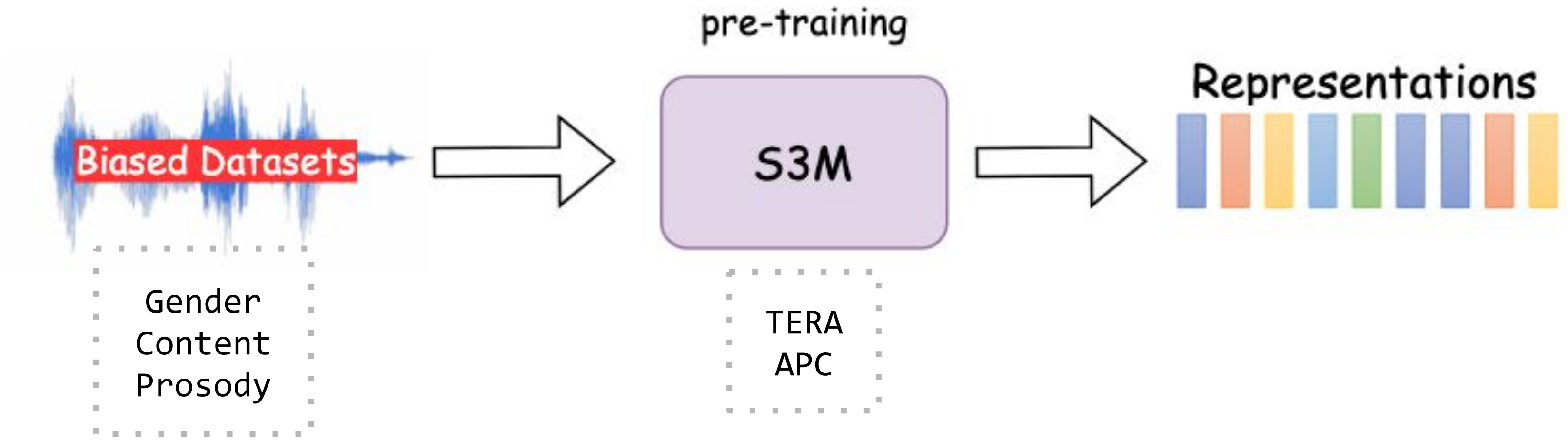


Outline

Phase 1

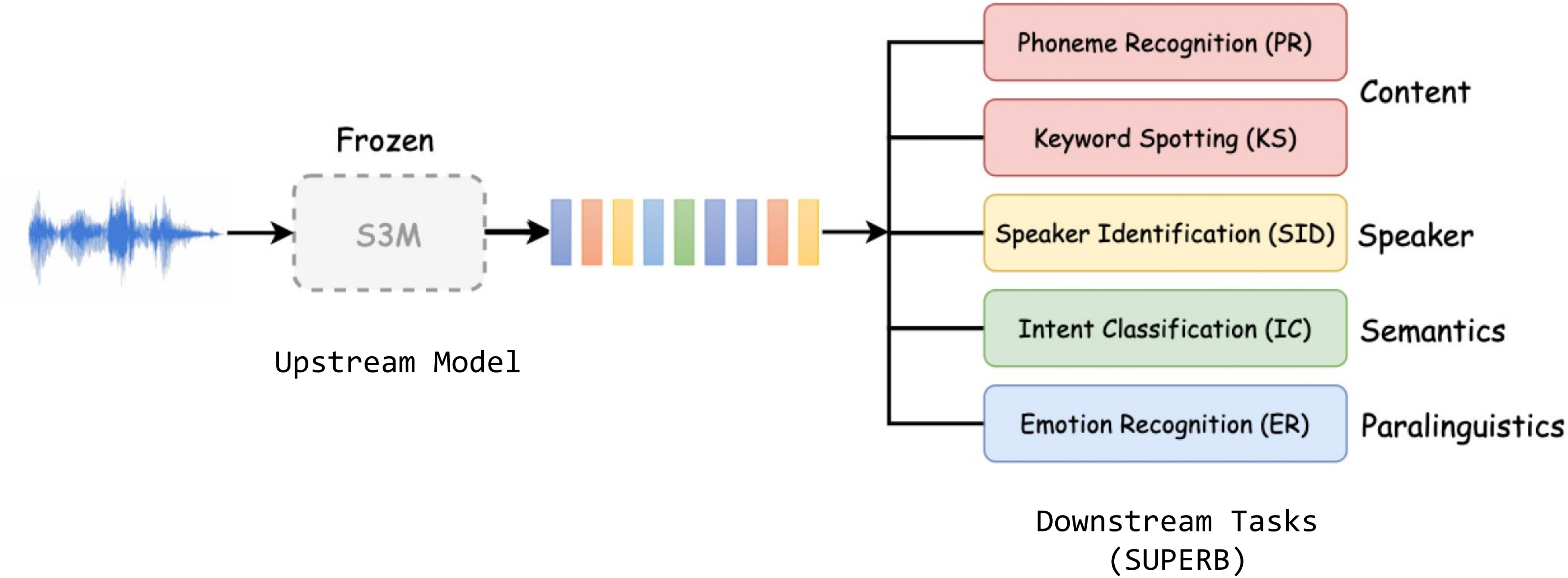
Self-supervised pre-training on biased datasets of these 3 aspects: **Gender, Content, and Prosody**

For the S3Ms, we select 2 models: **TERA and APC**



Phase 2

Evaluation on various downstream tasks from the **SUPERB** benchmark



Outline

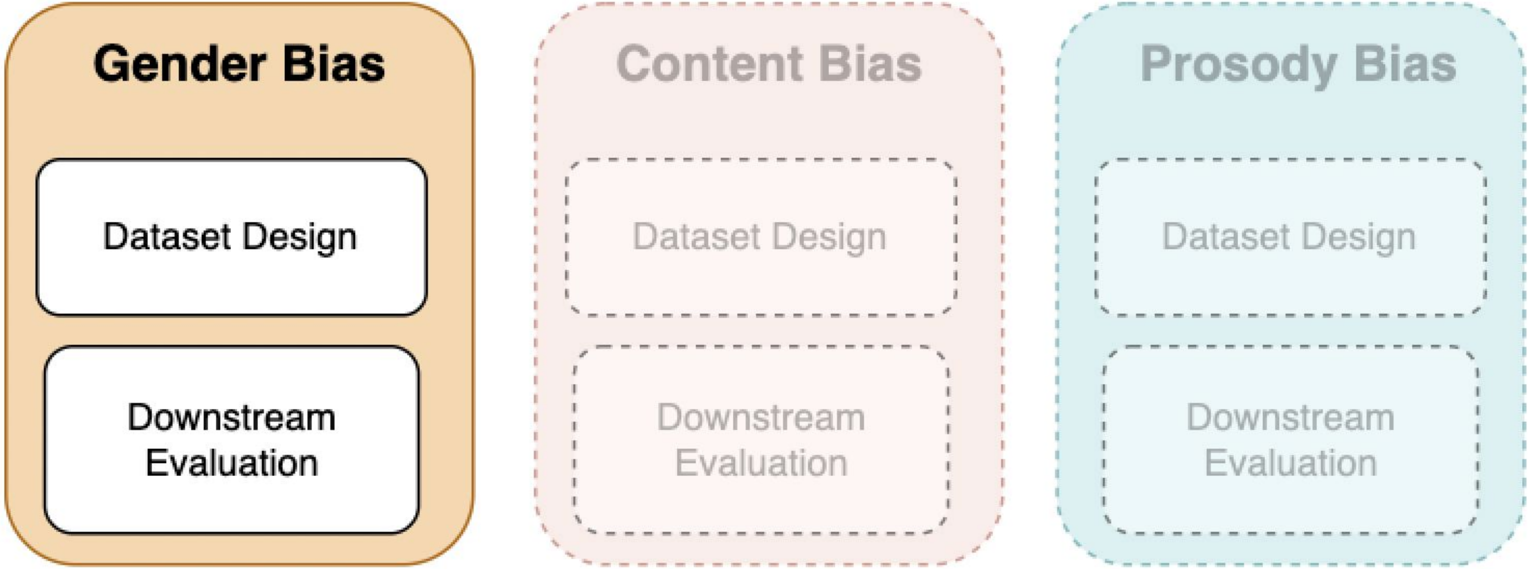
- Background & Motivation
- Introduction
- Methodology
- Experiments
- Conclusions

Pretraining Data

Pre-training data is fixed to 100 hours in all the experiments

LibriSpeech train-clean-100 (LS100) and LibriSpeech train-clean-360 (LS360) is used to design the 100-hr datasets with different biases

1. Gender


Gender Bias
Dataset Design
Downstream Evaluation
Content Bias
Dataset Design
Downstream Evaluation
Prosody Bias
Dataset Design
Downstream Evaluation

Dataset Design

Gender-biased Dataset:

3 random sampled datasets for each setting (3 x 6 = 18 datasets total):

- total 100 hrs from LS100 and LS360 with female-to-male ratio as 10:0, 9:1, 8:2, 2:8, 1:9, and 0:10 denote as All-F, 9F1M, 8F2M, 2F8M, 1F9M, and All-M

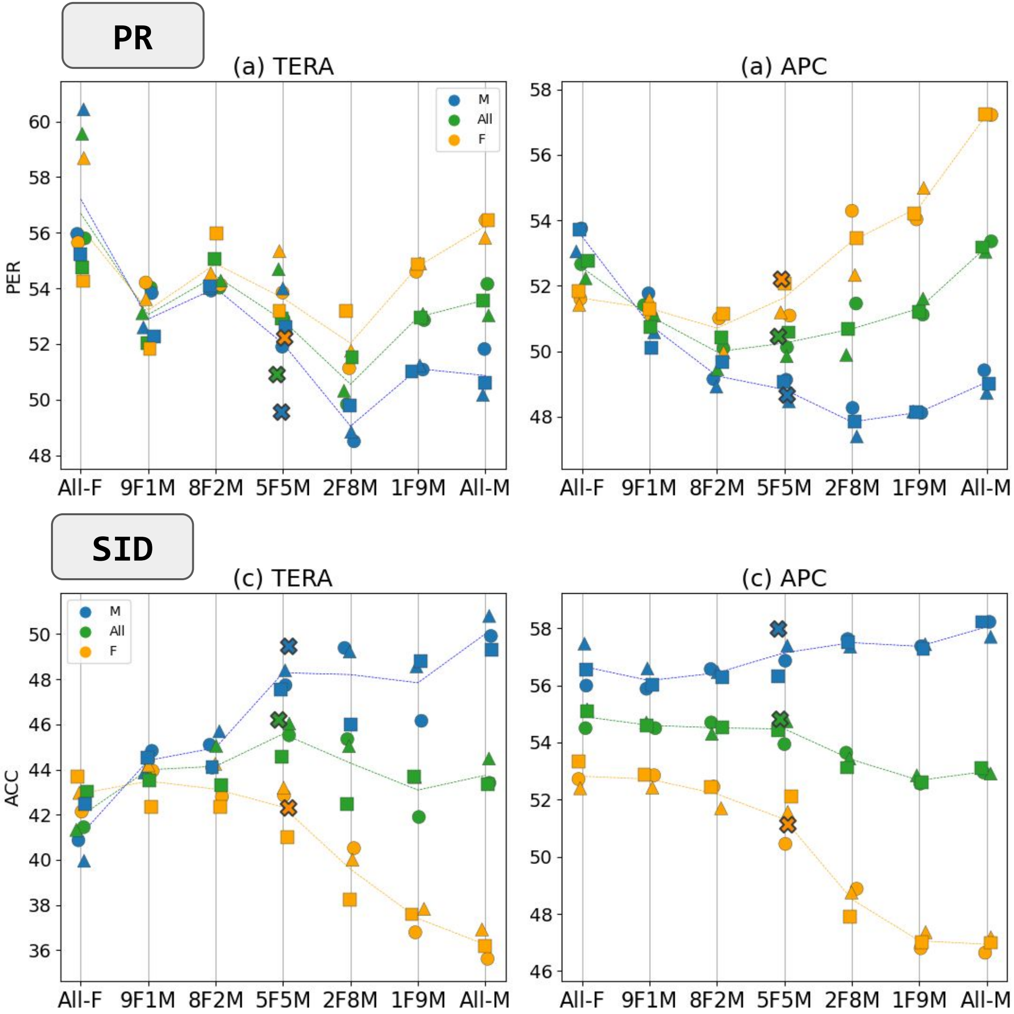
Baseline:

the original LS100 + 3 random sampled datasets with female-to-male ratio 5:5, denote as 5F5M

When testing on the downstream tasks, we spilt the testing data into **male and female subsets** if demographic information is provided

Downstream Evaluation

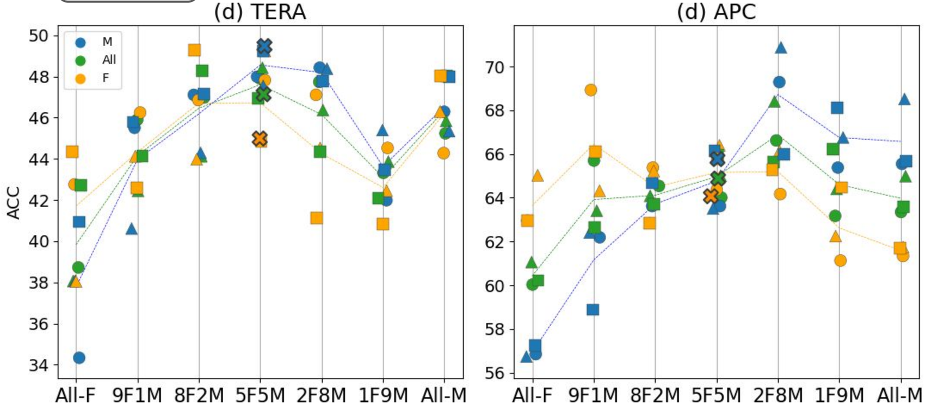
- Phoneme Recognition (PR) with APC and Speaker Identification (SID) with both S3Ms are more affected by gender bias
- Adding 10-20 percent of data can effectively bridge the gap between the testing accuracy of the male and female subsets



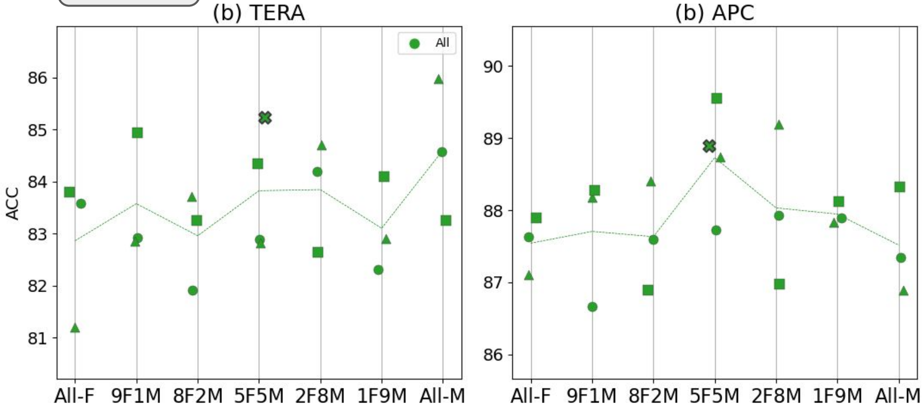
Downstream Evaluation

other tasks are comparatively irrelevant to gender bias

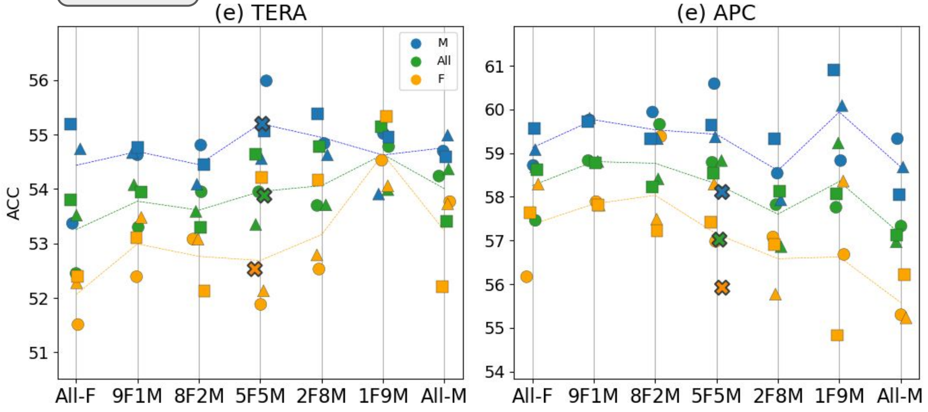
IC



KS



ER

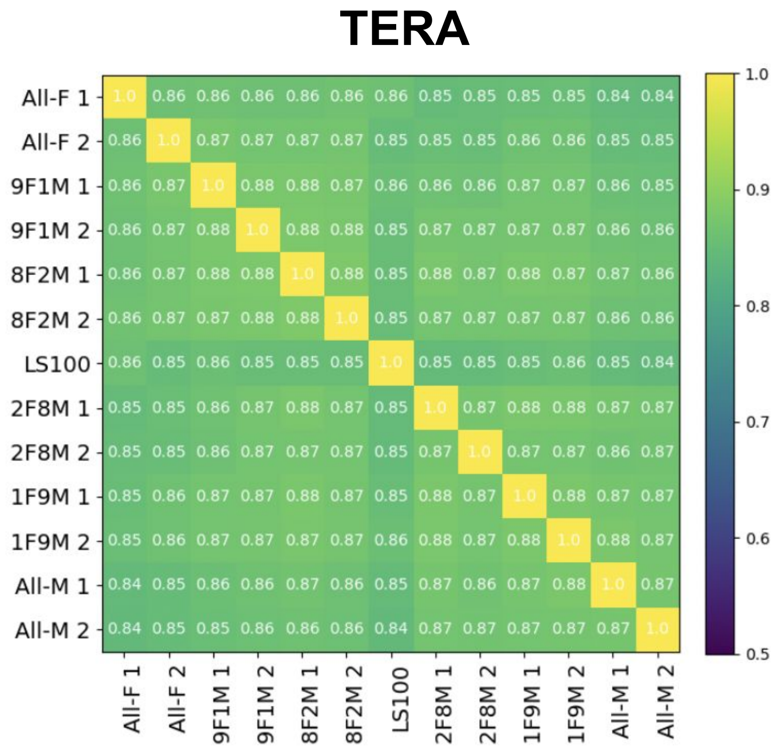


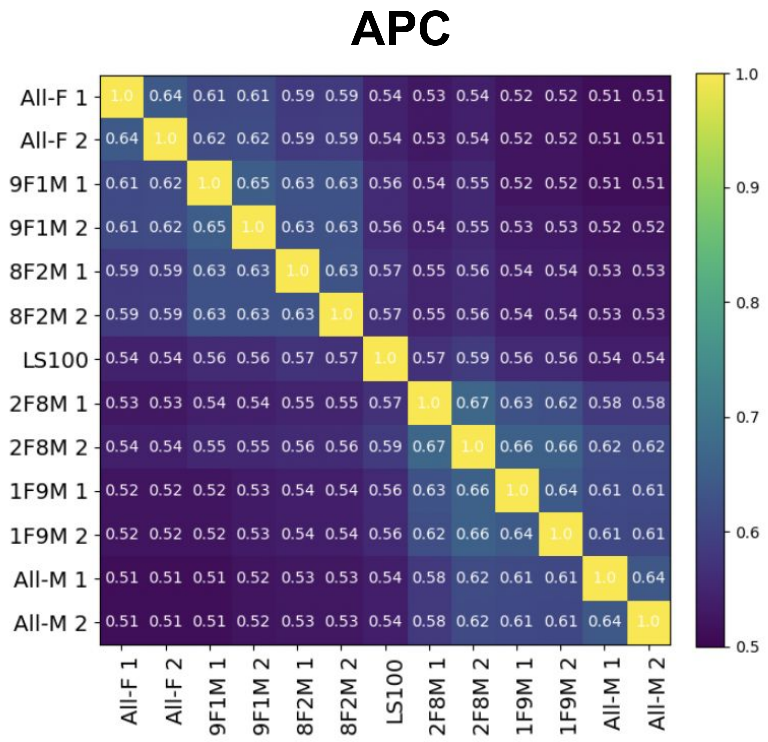
Representation Similarity

Further analysis of the representations extracted by the S3Ms pre-trained on different gender biased datasets

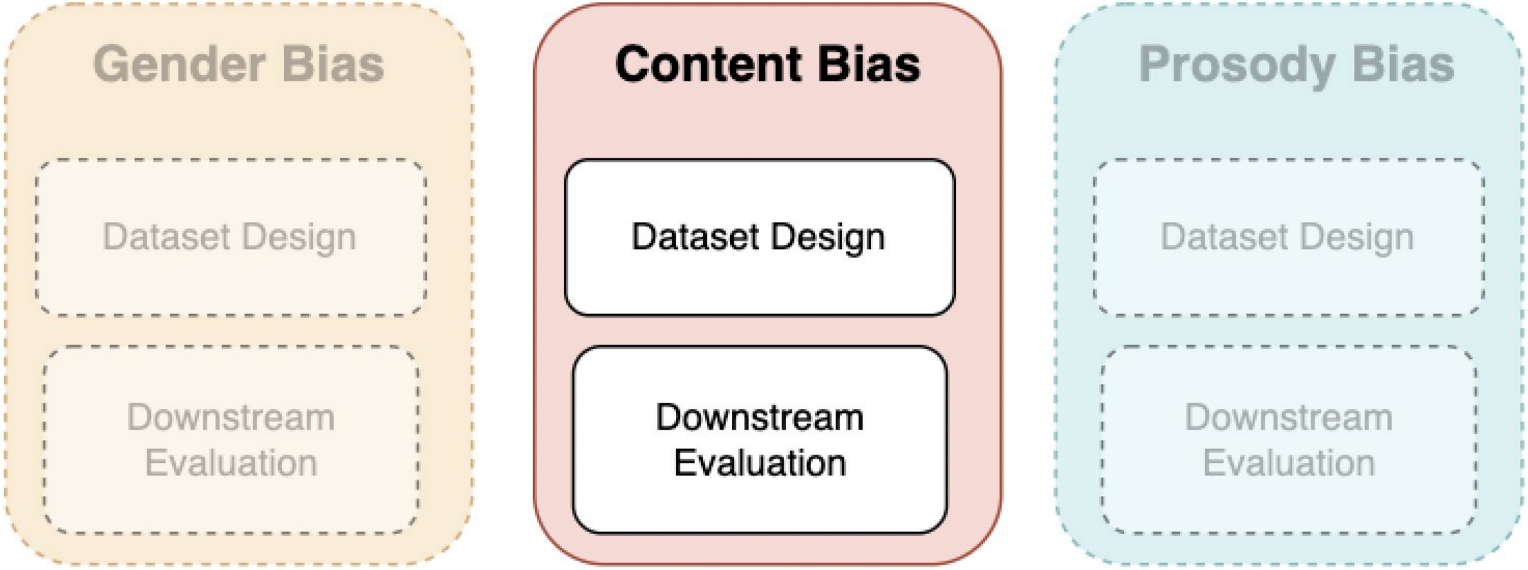
- Method: Projection Weighted Canonical Correlation Analysis (PWCCA)
- Dataset: LibriSpeech test-clean

No direct correlation between representation similarity and the behavior in downstream tasks





2. Content


Gender Bias
Dataset Design
Downstream Evaluation
Content Bias
Dataset Design
Downstream Evaluation
Prosody Bias
Dataset Design
Downstream Evaluation

Dataset Design

Content-biased Dataset:

calculate the perplexity (ppl) of the transcription of an utterance measured from the LS official ARPA language model

2 datasets:

- 100 hr audio with the highest ppl (ppl high)
- 100 hr audio with the lowest ppl (ppl low)

Baseline:

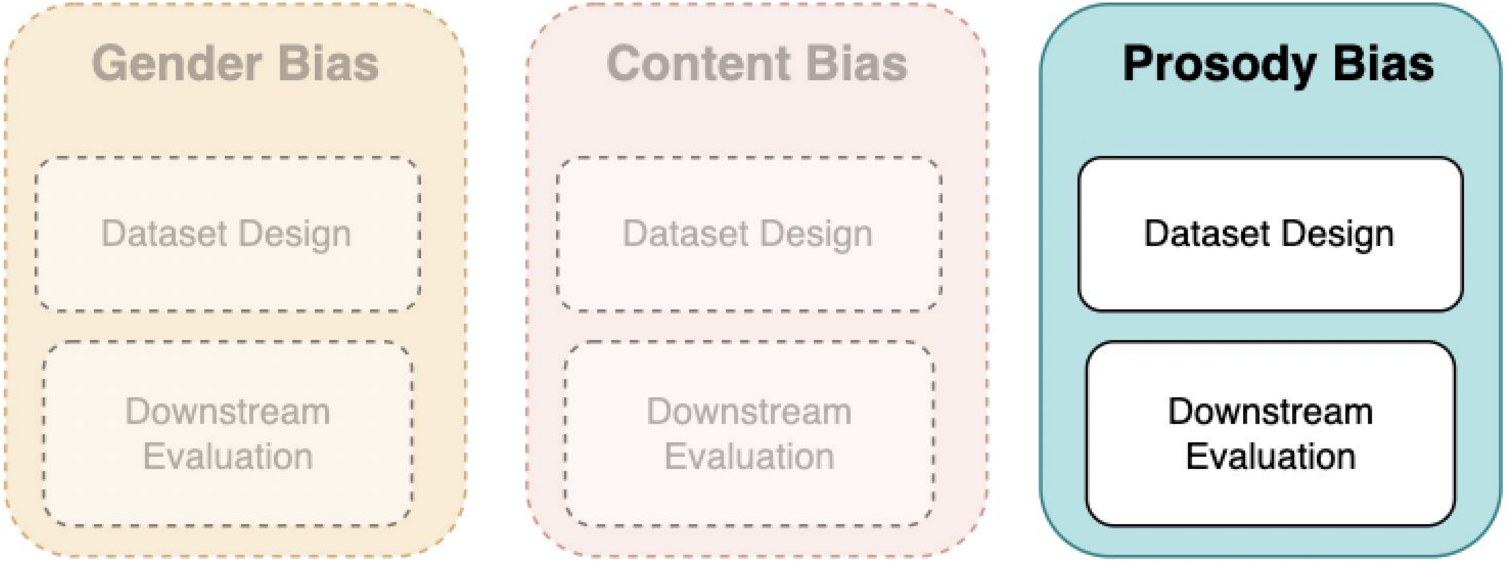
the original LS100

Downstream Evaluation

content bias would not affect the pre-training of S3Ms much

		PR PER ↓		KS ACC ↑		SID ACC ↑		IC ACC ↑		ER ACC ↑	
		TERA	**APC**	**TERA**	**APC**	**TERA**	**APC**	**TERA**	**APC**	**TERA**	**APC**
Baseline	LS 100	**49.64**	50.45	85.23	88.90	**46.20**	**56.94**	47.14	64.88	54.01	56.90
Content	ppl high	51.78	50.73	83.97	88.28	42.54	54.18	44.27	**65.67**	53.85	**58.46**
	ppl low	50.94	**50.17**	82.99	88.48	43.02	54.09	42.58	64.75	52.66	57.23
Prosody	wpm high	51.60	51.97	*81.37*	87.60	44.30	54.63	44.92	62.91	53.73	57.62
	wpm low	52.38	51.10	**86.37**	**89.13**	43.50	53.36	**49.93**	65.12	54.36	58.21
	speed 2x	*65.40*	*65.47*	81.73	*83.74*	*32.35*	*47.55*	*35.67*	*49.59*	*51.89*	*54.43*
	speed 0.5x	56.86	54.47	84.10	88.74	43.16	51.92	46.56	65.15	**54.43**	57.39

3. Prosody


Gender Bias
Dataset Design
Downstream Evaluation
Content Bias
Dataset Design
Downstream Evaluation
Prosody Bias
Dataset Design
Downstream Evaluation

Dataset Design

Prosody-biased Dataset:

4 datasets:

- relatively high/low speech rate: calculate words per minute(wpm) of each utterance
  - 100 hr audio with the highest wpm (wpm high)
  - 100 hr audio with the lowest wpm (wpm low)
- extreme speech rate:
  - convert the playbackspeed of LS100 two times faster (speed 2x)
  - convert the playbackspeed of LS100 two times slower (speed 0.5x)

Baseline:

the original LS100

Downstream Evaluation

Slower speech rate outperformed the baseline in some tasks

		PR PER ↓		KS ACC ↑		SID ACC ↑		IC ACC ↑		ER ACC ↑	
		TERA	**APC**	**TERA**	**APC**	**TERA**	**APC**	**TERA**	**APC**	**TERA**	**APC**
Baseline	LS 100	**49.64**	50.45	85.23	88.90	**46.20**	**56.94**	47.14	64.88	54.01	56.90
Content	ppl high	51.78	50.73	83.97	88.28	42.54	54.18	44.27	**65.67**	53.85	**58.46**
	ppl low	50.94	**50.17**	82.99	88.48	43.02	54.09	42.58	64.75	52.66	57.23
Prosody	wpm high	51.60	51.97	*81.37*	87.60	44.30	54.63	44.92	62.91	53.73	57.62
	wpm low	52.38	51.10	**86.37**	**89.13**	43.50	53.36	**49.93**	65.12	54.36	58.21
	speed 2x	*65.40*	*65.47*	81.73	*83.74*	*32.35*	*47.55*	*35.67*	*49.59*	*51.89*	*54.43*
	speed 0.5x	56.86	54.47	84.10	88.74	43.16	51.92	46.56	65.15	**54.43**	57.39

Outline

- Background & Motivation
- Introduction
- Methodology
- Experiments
- Conclusions

Conclusions

- Pre-training data does not need to be gender-balanced to ensure the best performance
- Content bias in pre-training data does not affect much
- S3Ms show a preference towards slower speech rate