Paper ID: 8811

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

Introduction

Background

- In real-world applications, collected audio data can be biased in different aspects.

Demographic: gender, age, accent, …

Content: topic, word use, …

Prosody: speech rate, tone, …

- Self-Supervised Speech Models(S3Ms) are often pre-trained on "standard" datasets such as LibriSpeech

→ The effect of data bias in S3Ms is unexplored

Research Question

- How would data bias in S3Ms pre-training affect downstream tasks?

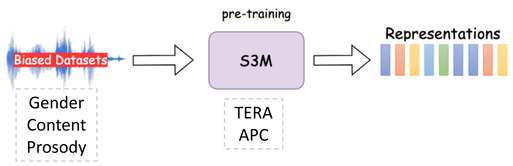
Key Observation

- 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 a slower speech rate

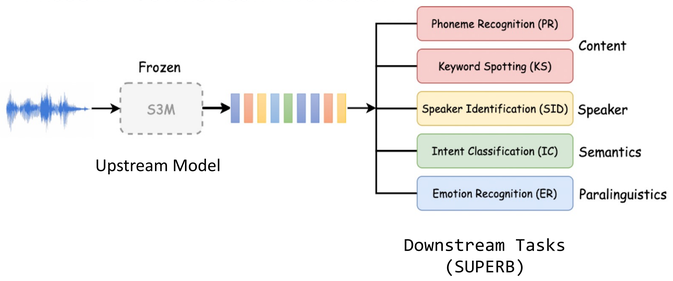
Framework

Phase 1 - Self-Supervised Pre-training

- All datasets are fixed to **100 hours**



Phase 2 - Downstream Evaluation



Experiment Setup

Upstream Models (S3Ms)

TERA: Self-Supervised Learning of Transformer Encoder Representation for Speech

- Inspired by masked language modeling(MLM)
- Learning by predicting altered frames

APC: An Unsupervised Autoregressive Model for Speech Representation Learning

- Inspired by language models (LMs) for text
- Learning to predict future frames

Downstream Tasks

5 selected tasks from the SUPERB benchmark, categorized into 4 aspects

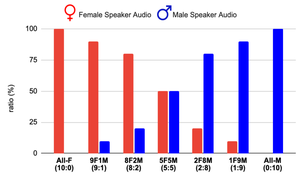
- **Content:** Phoneme Recognition (PR), Keyword Spotting (KS)
- **Speaker:** Speaker Identification (SID)
- **Semantics:** Intent Classification (IC)
- **Paralinguistics:** Emotion Recognition (ER)

Results

Gender Bias

Biased Datasets (6 settings x 3 random = total 18)

- All-F, 9F1M, 8F2M, 2F8M, 1F9M, All-M



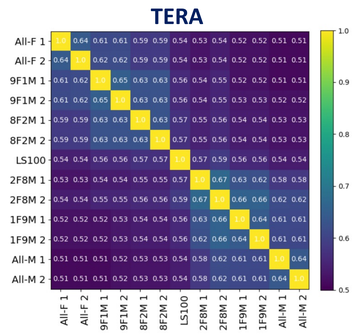
Baseline Datasets (total 4)

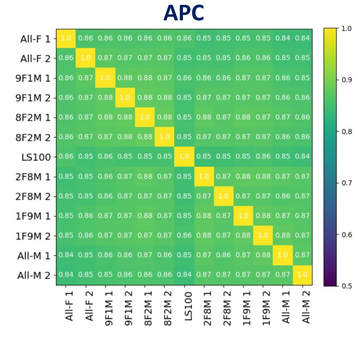
- The original LS100 + 3 random sampled 5F5M datasets

Representation Similarity Heatmap

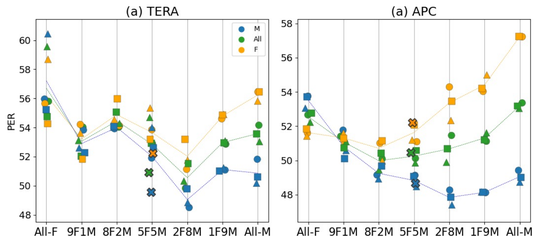
The similarity score between each pair of models pre-trained on different *gender*-biased datasets.

- Projection Weighted Canonical Correlation Analysis (PWCCA)
- Mean similarity score of the utterances in LibriSpeech *test-clean*

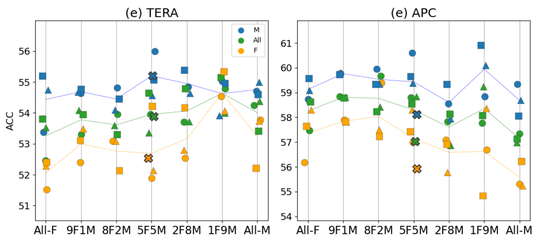




Phoneme Recognition(PR): ***More Affected by Gender Bias***



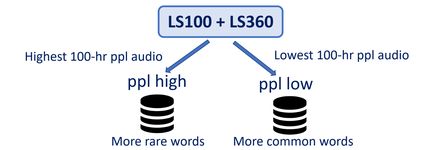
Intent Classification(IC): ***Comparatively Irrelevant to Gender Bias***

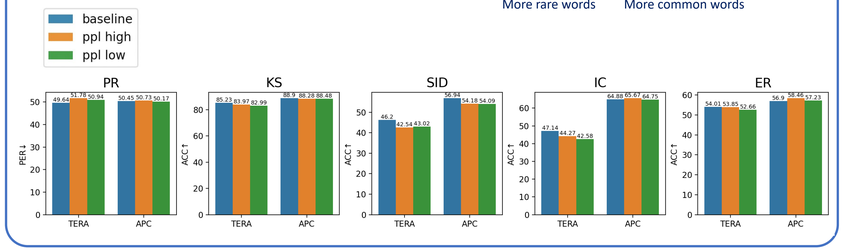


Content Bias

Biased Datasets (total 2):
Biased in word use

- Measurement: calculate the perplexity (ppl) of each utterance measured from the LS official LM

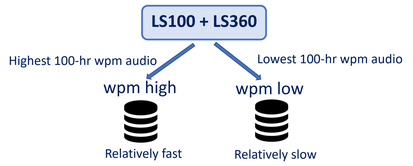




Prosody Bias

Biased Datasets (total 4):
Focus on the speech rate difference

Relatively biased speech rate
Measurement: calculate words per minute(wpm) for each utterance



Extremely biased speech rate
Convert the playback speed of each audio in LS100 2 times faster/slower