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*[⋆†]

[⋆]Graduate Institute of Communication Engineering, National Taiwan University
[†]College of Electrical Engineering and Computer Science, National Taiwan University
{r10942085, b06901012, f07942089, hungyilee}@ntu.edu.tw

ABSTRACT

Self-supervised Speech Models (S3Ms) have been successful in many speech downstream tasks, like ASR. However, how pre-training data affects S3Ms' downstream behavior remains an unexplored issue. In this paper, we study how pre-training data affects S3Ms by pre-training models on biased datasets targeting different factors of speech, including gender, content, and prosody, and evaluate these pre-trained S3Ms on selected downstream tasks in SUPERB Benchmark. Our experiments show that S3Ms have tolerance toward gender bias. Moreover, we find that the content of speech has little impact on the performance of S3Ms across downstream tasks, but S3Ms do show a preference toward a slower speech rate.

Index Terms— Self-supervised Speech Models, SUPERB Benchmark, Data Bias

1. INTRODUCTION

Self-supervised Learning (SSL) from raw speech has become increasingly popular in recent studies, as SSL achieves state-of-the-art results on various downstream tasks [1–3], ranging from speaker identification, automatic speech recognition, intent classification and a lot more. Recent work also dive into the interpretability of Self-supervised Speech Models (S3Ms), as models can be sensitive to powerful adversarial attacks. Researchers are curious about what these models have really learned, and most work focus on the explainability of model mechanisms or learned representations [4–6]. Nevertheless, how pre-training data affects S3Ms is less studied.

Self-supervised pre-training in most works use only standard datasets, for example, LibriSpeech [7], with carefully collected, gender-balanced, and clean data. While it is easy to collect a large amount of unlabeled data, collecting "balanced" data is relatively hard in practical applications. Besides the most discussed gender bias, speech data can also be biased in other aspects, such as content and prosody. However, to our best knowledge, how bias in speech affects SSL is yet unknown.

In this paper, we present an empirical study on the effects of data bias on S3Ms at the pre-training stage by creating various datasets. The performance of S3Ms on different pre-training data is evaluated by selected tasks from SUPERB Benchmark [8]. Our study in three aspects of data bias provides the following insights :

1. **Gender**: We pre-train S3Ms on datasets with different gender distributions. Our study shows that gender-balanced data is not necessarily needed for effective pre-training.

2. **Content**: We pre-train S3Ms on two groups of data with extremely biased speech content. We find that content has little effect on the downstream behavior of S3Ms.

3. **Prosody**: We pre-train S3Ms on datasets with faster or slower speech rates. Experiments show that S3Ms pre-trained on slower speech rate lead to better performance overall.

2. RELATED WORK

One of the most often used learning schemes for self-supervised speech models is through reconstruction speech frames. Here we introduce some of the recently emerged reconstruction methods. The Autoregressive Predictive Coding (APC) method [9], is primarily inspired by language models (LM) for text. The DeCoAR [10] method combines the bidirectionality of ELMo [11] and the autoregressive reconstruction objective of APC [9]. The work of [12–14] also adopt the autoregressive reconstruction scheme, but in some variation. The Transformer Encoder Representations from Alteration (TERA) [15] method, an improved version of Mockingjay [16], is mainly inspired by masked language models (MLM) [17] for text. The work of [18–20] also adopt variations of the MLM reconstruction schemes. In this work, we select two methods to represent each scheme for our study, the APC method from the autoregressive family and the TERA method from the MLM family. We select two models for our study due to space limitations.

Bias and fairness issues in speech are receiving more attention these days. Demographic bias is the most studied. Recent works analyze the impact of demographic bias as well as mitigating bias on specific tasks including ASR [21, 22], speaker recognition [23, 24]. and speech translation [25–27]. A large body of research related to data bias evaluates models on a single downstream task, however, data bias on S3Ms and its effect on downstream tasks from diverse categories are not yet explored. A related work analyzing pre-training data of S3Ms is [28], which investigates the effect of domain shift in SSL. Our work differs from theirs in the sense that we focus on data bias toward different speech factors at pre-training stage.

3. EXPERIMENTAL SETUP

As we attempt to investigate the impact of pre-training data bias on S3Ms, the settings for fine-tuning, including fine-tuning data and hyperparameters, are all the same, the only difference is pre-training data.

---

*Equal Contribution

3.1. Pre-trained Upstream Models

For our experiments, we consider two of the most representative S3Ms, Transformer Encoder Representations from Alteration (TERA) [15] and Autoregressive predictive coding (APC) [9].

- **TERA:** As suggested by the TERA paper, we use two of the alterations proposed by the authors: time and frequency, and pre-train models for 200,000 steps with a batch size of 32.

- **APC:** As suggested by the APC paper, we train APC for 100 epochs with a batch size of 32 and use ADAM optimizer with an initial learning rate of 10^{-4}.

3.2. Datasets

We hope to explore how bias toward gender distribution, content, and prosody can affect S3Ms, therefore, we design various artificial datasets for pre-training and further evaluate models on downstream tasks. For faster pre-training and fairness settings, we fix our pre-training data to 100 hr in all the experiments. We use LibriSpeech (LS) 100 hr and 360 hr to design the 100-hr datasets with different biases as below.

3.2.1. Gender

Here, We pay attention to the behavior of S3Ms when pre-training data is biased toward gender distribution. Thus, we design datasets with male-to-female ratio as 0:10, 1:9, 2:8, 8:2, 9:1, and 10:0 by randomly sampling files from LS 100 hr and 360 hr. These 6 settings are denoted as *All-F*, *9F1M*, *8F2M*, *2F8M*, *1F9M*, and *All-M* respectively. To better interpret the results, we randomly sample three 100-hr datasets for each of the gender distribution settings above. For 5:5 male to female ratio (denoted as *5F5M*), we use the original LS 100 subset[1] as well as 3 random sampled 100-hr datasets from LS 100 hr plus 360 hr.

3.2.2. Content

In this section, we aim to explore whether pre-training on "complex" or "simple" sentences affects S3Ms' downstream behavior. Here we use the perplexity (ppl) of the transcription of an utterance measured from a language model to determine whether a sentence is complex or simple. We utilize the LS official ARPA language model to calculate ppl for each transcription in the LS 100 hr and 360 hr subset and create two datasets, 100 hr audio with the highest ppl and 100 hr audio with the lowest ppl, denoted as *ppl high* and *ppl low* respectively. Audios in *ppl high* contain rarer words and proper nouns, while audios in *ppl low* are mostly composed of common and simple words.

3.2.3. Prosody

In addition to gender and content, prosody is also an essential aspect of speech study. Speech rate is viewed as an important prosodic feature, hence in this section, we design the datasets based on speech rates as below. We calculate words per minute (wpm) for each sentence in the LS 100 hr and 360 hr subset using the alignments of utterances and the provided transcriptions. Similar to the setup in Section 3.2.2, we create two datasets, 100 hr audio with the highest wpm and 100 hr audio with the lowest wpm, denoted as *wpm high* and *wpm low* respectively. Moreover, to further investigate the

---

[1]The original LS 100 subset is gender balanced.

---

impact of extreme speech rates on S3Ms, we create two additional artificial datasets by converting the playback speed of all audio files in LS 100 hr subset two times faster and two times slower without altering the pitch. These two datasets are denoted as *speed 2x* and *speed 0.5x* respectively.

3.3. Downstream Tasks

To evaluate the generalizability and effectiveness of pre-trained models on diverse tasks, we select five tasks, which are solvable with linear downstream models, from SUPERB Benchmark [8]. Tasks are carefully chosen as we wish to analyze the effect of data bias, and linear models serve as the direct indication of the quality of speech representations.

For the selected tasks, we follow the settings in the SUPERB Benchmark. The only difference is that instead of using weighted sum to integrate hidden states from all layers, we directly use the last hidden state learned from S3Ms. Salient performance gap may occur for S3Ms with a larger architecture, but since there are only three layers in both TERA and APC, we observe no huge difference between features with and without weighted sum on downstream performance. The five tasks can be further categorized into four aspects of speech: content, speaker, semantics, and paralinguistics:

|	
Content	**Phoneme Recognition, PR**: LibriSpeech [7] is adopted. The evaluation metric is phone error rate(PER).
	Keyword Spotting, KS: Speech Commands dataset v1.0 [29] is adopted. The evaluation metric is accuracy(ACC).
Speaker	**Speaker Identification, SID**: VoxCeleb1 [30] is adopted. The evaluation metric is accuracy(ACC).
Semantics	**Intent Classification, IC**: Fluent Speech Commands dataset [31] is adopted. The evaluation metric is accuracy(ACC).
Paralinguistics	**Emotion Recognition, ER**: IEMOCAP [32] is adopted. The evaluation metric is accuracy(ACC).

4. RESULTS AND ANALYSIS

4.1. Gender

4.1.1. Downstream Behavior

Figure 1 shows that, in general, S3Ms pre-trained on balanced data achieve the best result; however, pre-training models on gender-imbalanced data does not always cause a severe degrade in downstream performance. At the most extreme setting, namely *All-F* and *All-M*, most of S3Ms perform the worst. But this situation can be mitigated by adding a small amount of data from the other gender, which can effectively elevate the performance, models can then be on par with, or even better than, those pre-trained on balanced data.

The testing sets of selected tasks are approximately gender-balanced. For further investigation on gender bias during testing, we split the testing set by gender and evaluate S3Ms on male and female subsets separately, except for the dataset of KS, where no
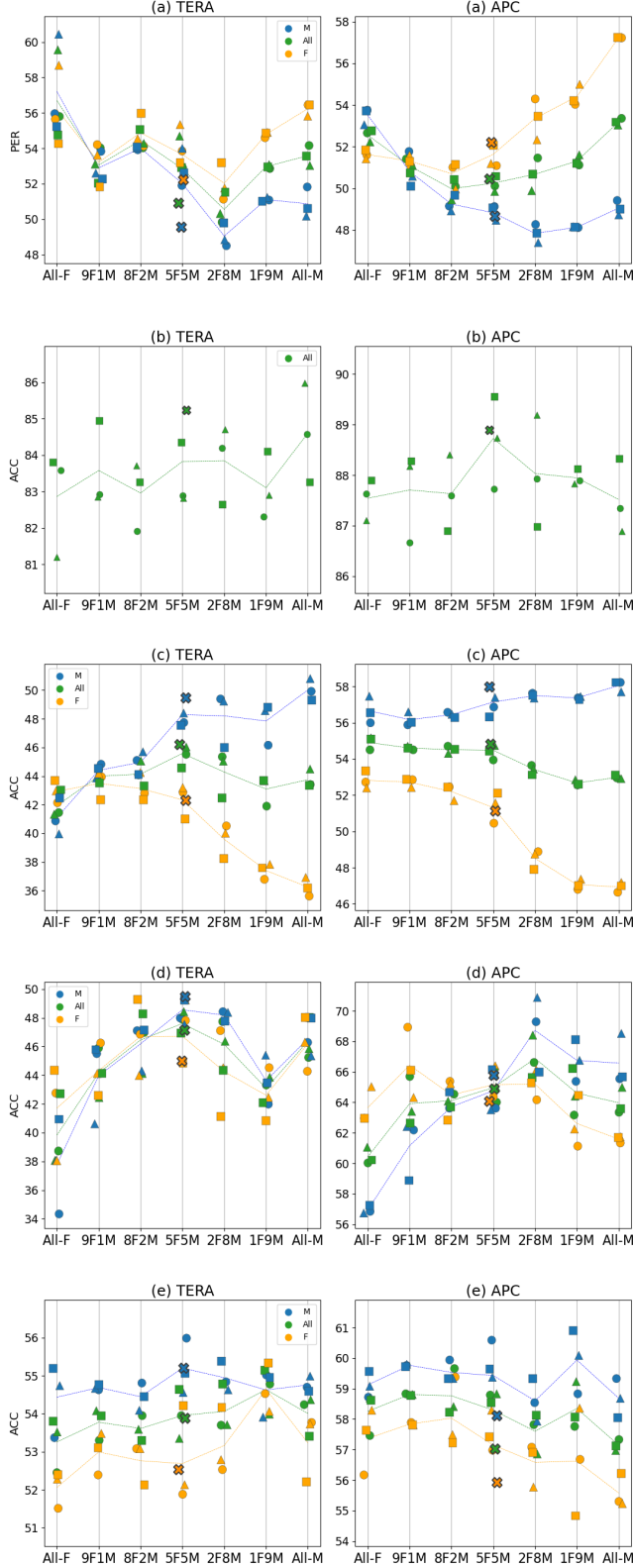
Fig. 1: **(a) PR (b) KS (c) SID (d) IC (e) ER**
Results (in %) of both S3Ms (TERA and APC) pre-trained on data with different gender distribution. Data points come with three different shapes, indicating three random sampled pre-training dataset. The dashed line is obtained by connecting the average of three data points. In 5F5M, the 'x' marker with bolder outline represents the LS 100 hr dataset. Notation—'M': `male`, 'F': `female`

demographic information is provided. As such, the overall testing score will be the average of scores on male and female subsets.

We observe that some tasks seem to be affected more by gender bias, such as PR with APC and SID with both S3Ms. When pre-training models on datasets with higher male voice ratio, the accuracy on female subset drops rapidly, and vice versa. We conjecture this is because the diversity of female voice is much higher than that of male voice. However, we can bridge the gap between the testing accuracy of the male and female subsets by just adding 10-20 percent of female voice. For other tasks, gender bias is not obviously presented, for instance, ER is comparatively agnostic to gender bias.

Moreover, figure 1 shows that the behaviors of the models differ across downstream tasks. In comparison of the two models, the connected mean line in APC is smoother, while in TERA, the randomness in three random sampled datasets is higher. Still, the overall performance trend in the selected tasks is fairly alike.

4.1.2. Representation Similarity

Following our results in section 4.1.1, we find that the effect on downstream tasks is not as significant as we originally expected, even when gender is highly imbalanced in pre-training data. Hence we are curious whether the representations extracted from S3Ms pre-trained on different gender-biased datasets are similar. Therefore, we measure the representation similarity of S3Ms pre-trained on gender-biased datasets. For similarity measurement, we adopt Projection Weighted Canonical Correlation Analysis (PWCCA) proposed in [33], and we use LibriSpeech *test-clean* subset.

From figure 2 and figure 3, TERA and APC behave differently as we take a closer look at the PWCCA score of their upstream representations. For TERA, the representations of different gender-biased datasets are all very similar, yet we cannot see a higher similarity between two random sampled datasets with the same gender distribution. For APC, the overall similarity between different datasets is much lower than that in TERA. However, the upper left corner block and the bottom right corner block show a lighter color, meaning that the representation similarity increases when the gender distribution is more similar. For example, the representation of All-F is much more similar to 9F1M than 1F9M. Also, the similarity is the highest between different random sampled datasets under the same gender distribution setting.

While similarity scores for TERA with different gender bias settings are highly alike, we can observe a correlation between gender distribution and its similarity scores for APC. However in both S3Ms, gender-bias in pre-training data has only a slight effect in downstream. As a result, there is no obvious relationship between representation similarity and small gender bias in downstream tasks.

4.2. Content

Table 1 lists the testing results of S3Ms (TERA and APC) pre-trained on content and prosody bias. Surprisingly, we observe that, for both TERA and APC, there is little performance difference between pre-training on either content-biased dataset (*ppl high* and *ppl low*), even evaluated on tasks related to content. For TERA, there is a slight performance drop across five downstream tasks. But for APC, pre-training on content-biased data barely degrades the testing results, and models even outperform the baseline on some tasks such as ER and IC.

	pre-train	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

Table 1: The testing result (in %) of S3Ms pre-trained on designed datasets. The arrow in the header indicates whether lower/higher score is better. The bold text denotes the best performance on the column, and the red text denotes the worst performance on the column. *Note that for each task, the performance is only compared with the same model.*
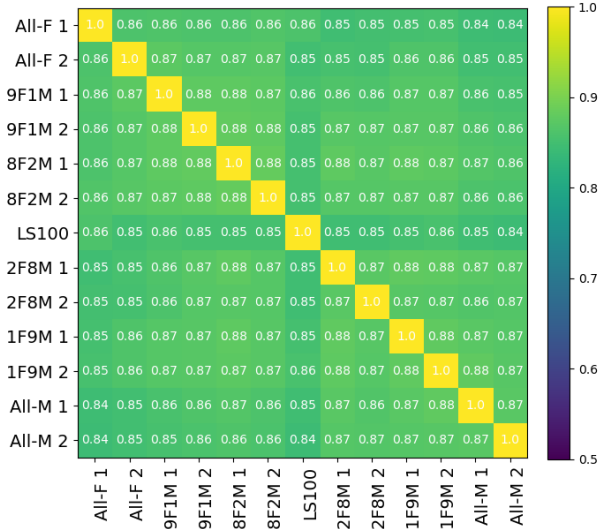
Fig. 2: Similarity heatmap among different gender setting in **TERA**. Similarity values are annotated. The number following gender setting indicates different random sampled pre-training data
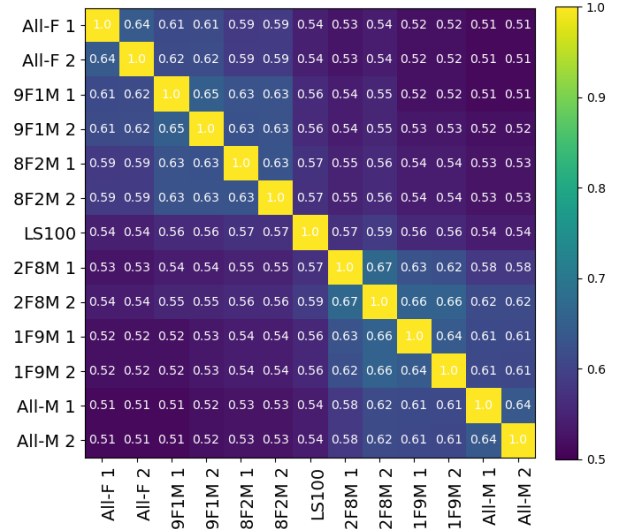
Fig. 3: Similarity heatmap among different gender setting in **APC**. Similarity values are annotated.

4.3. Prosody

As table 1 shows, the performance difference between *wpm high* and *wpm low* on PR, SID, and ER is not obvious. Nevertheless, we see that data with slower speech rate performs significantly better on KS and IC, especially in TERA.

For *speed 2x* and *speed 0.5x*, a significant difference in downstream performance can be observed. Pre-training on *speed 2x* has a considerable performance drop across all tasks. While using *speed 0.5x* for pre-training slightly degrades performance on PR, KS, and SID, the performance is on par with baseline in terms of IC and even achieves better results on ER. Interestingly, pre-training S3Ms on data with lower speech rate performs better than data with higher speech rate — 6 out of 10 even outperform the baseline (LS 100). Results suggest that it is not too harmful to pre-train on data with extremely slow speech rate, instead, slower speech rate may even be beneficial to some tasks.

5. CONCLUSIONS

Our work presents an empirical approach for understanding the effect of biased data on S3Ms. The quality of speech representations affected by data bias is carefully examined, as we evaluate S3Ms on a wide range of downstream tasks with linear models. Aside from downstream modeling, we also measure representation similarity for more insights, and results show no direct correlation between downstream behavior and representation similarity. Results on gender bias show that pre-training data does not need to be gender-balanced to ensure the best performance on downstream tasks. Furthermore, our study suggests that pre-training on biased content does not affect much. Finally, we find that pre-training S3Ms on data with lower speech rate achieves better performance. For future work, the effect of data bias can be studied on more S3Ms from different families. We are also interested to explore other aspects of bias, for instance, single/multiple speakers, synthesized/natural speech, and noisy/quiet environments.

6. REFERENCES

[1] Ha Nguyen, Fethi Bougares, N. Tomashenko, Yannick Estève, and Laurent Besacier, "Investigating Self-Supervised Pre-Training for End-to-End Speech Translation," in *Proc. Interspeech*, 2020.

[2] Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: How much can a bad teacher benefit asr pre-training?," in *ICASSP*, 2021.

[3] Sung-Feng Huang, Shun-Po Chuang, Da-Rong Liu, Yi-Chen Chen, Gene-Ping Yang, and Hung yi Lee, "Self-supervised pre-training reduces label permutation instability of speech separation," *ArXiv*, vol. abs/2010.15366, 2020.

[4] Sercan Ö. Arik and Tomas Pfister, "Tabnet: Attentive interpretable tabular learning," *Proc. AAAI Conference on Artificial Intelligence*, 2021.

[5] Shu wen Yang, Andy T. Liu, and Hung yi Lee, "Understanding self-attention of self-supervised audio transformers," 2020.

[6] Yu-An Chung, Yonatan Belinkov, and James Glass, "Similarity analysis of self-supervised speech representations," in *ICASSP*, 2021.

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio book," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2015.

[8] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021.

[9] Yu-An Chung and James Glass, "Generative pre-training for speech with autoregressive predictive coding," in *ICASSP*, 2020.

[10] S. Ling, Y. Liu, J. Salazar, and K. Kirchhoff, "Deep contextualized acoustic representations for semi-supervised speech recognition," in *ICASSP*, 2020.

[11] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, "Deep contextualized word representations," *NAACL-HLT*, 2018.

[12] Y. Chung and J. Glass, "Generative pre-training for speech with autoregressive predictive coding," in *ICASSP*, 2020.

[13] Yu-An Chung, Hao Tang, and James Glass, "Vector-quantized autoregressive predictive coding," *Interspeech*, 2020.

[14] Yu-An Chung and James Glass, "Improved speech representations with multi-target autoregressive predictive coding," in *Proc. ACL*, 2020.

[15] Andy T. Liu, Shang-Wen Li, and Hung yi Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," 2020.

[16] Andy T. Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," *ICASSP 2020*, May 2020.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[18] Song Li, Lin Li, Qingyang Hong, and Lingling Liu, "Improving Transformer-Based Speech Recognition with Unsupervised Pre-Training and Multi-Task Semantic Knowledge Learning," in *Interspeech*, 2020.

[19] Dongwei Jiang, Wubo Li, Ruixiong Zhang, Miao Cao, Ne Luo, Yang Han, Wei Zou, Kun Han, and Xiangang Li, "A further study of unsupervised pretraining for transformer based speech recognition," in *ICASSP*, 2021.

[20] Lu Liu and Yiheng Huang, "Masked pre-trained encoder base on joint ctc-transformer," 2020.

[21] Josh Meyer, Lindy Rauchenstein, Joshua D. Eisenberg, and Nicholas Howell, "Artie bias corpus: An open dataset for detecting demographic bias in speech applications," in *LREC*. May 2020, European Language Resources Association.

[22] Joshua L Martin, "Spoken corpora data, automatic speech recognition, and bias against african american language: The case of habitual 'be'," in *Proc. ACM FAccT*, 2021.

[23] Gianni Fenu, Mirko Marras, Giacomo Medda, and Giacomo Meloni, "Fair Voice Biometrics: Impact of Demographic Imbalance on Group Fairness in Speaker Recognition," in *Proc. Interspeech*, 2021.

[24] Gianni Fenu, Giacomo Medda, Mirko Marras, and Giacomo Meloni, "Improving fairness in speaker recognition," 2021.

[25] Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia Antonino Di Gangi, Roldano Cattoni, and Marco Turchi, "Gender in danger? evaluating speech translation technology on the must-she corpus," 2020.

[26] Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi, "Breeding gender-aware direct speech translation systems," 2020.

[27] Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi, "How to split: the effect of word segmentation on gender bias in speech translation," 2021.

[28] Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," 2021.

[29] Pete Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 2018.

[30] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, 2020.

[31] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio, "Speech model pre-training for end-to-end spoken language understanding," 2019.

[32] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, 2008.

[33] Ari Morcos, Maithra Raghu, and Samy Bengio, "Insights on representational similarity in neural networks with canonical correlation," in *NeurIPS*. 2018.