Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6462–6468
Marseille, 11–16 May 2020

Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications

Josh Meyer,[†] **Lynn Rauchenstein,**[†] **Joshua D. Eisenberg,**[†] **Nicholas Howell** [◦]
[†] Artie, Inc. [◦] Higher School of Economics
Los Angeles, CA, USA Moscow, Russian Federation
{josh.meyer, lynn.rauchenstein, joshua.eisenberg}@artie.com, nlhowell@gmail.com

Abstract

We describe the creation of the Artie Bias Corpus, an English dataset of expert-validated **<audio, transcript>** pairs with demographic tags for **{age, gender, accent}**. We also release open software which may be used with the Artie Bias Corpus to detect demographic bias in Automatic Speech Recognition systems and can be extended to other speech technologies. The Artie Bias Corpus is a curated subset of the Mozilla Common Voice corpus, which we release under a Creative Commons CC-0 license – the most open and permissive license for data. This article contains information on the criteria used to select and annotate the Artie Bias Corpus in addition to experiments in which we detect and attempt to mitigate bias in end-to-end speech recognition models. We observe a significant accent bias in our baseline DeepSpeech model, with more accurate transcriptions of US English compared to Indian English. We do not, however, find evidence for a significant gender bias. We then show significant improvements on individual demographic groups from fine-tuning.

Keywords: speech corpus, automatic speech recognition, demographic bias, bias detection

1. Introduction

Speech technologies such as Automatic Speech Recognition (ASR) and Speaker Verification have become an everyday part of life in many countries. However, as the technology becomes more common we are discovering how fragile it can be. Environmental noise is a clear source of performance degradation, but other (less visible) sources of degradation stem from demographic factors such as age, gender, and accent (Hashimoto et al., 2018; Garnerin et al., 2019; Tatman, 2017; Tatman and Kasten, 2017). A speech technology exhibits demographic bias when performance is worse for one demographic group relative to another.

Demographic sources of bias in speech technology are usually the result of imbalanced training datasets. Most research and development in Automatic Speech Recognition for English has been performed on datasets comprised of a majority of white American English speakers. Even though this is a known problem, to the best of our knowledge there exist no free and open resources to diagnose (in an attempt to mitigate) this bias. The current lack of resources is the motivation behind the Artie Bias Corpus[1] and associated tools[2].

2. Prior work

2.1. Bias detection

Previous work in the detection of demographic bias in Speech Recognition models can be found in Tatman (2017) and Tatman and Kasten (2017). In Tatman (2017) the author studies performance of the proprietary ASR systems used in Youtube automated captioning on a collection of "accent tag challenge" videos. The corpus used to identify bias was a crowd-sourced set of speakers self-identifying their accent and then reading a list of isolated words known to have high variation in pronunciation. This corpus was relatively small (62 words read in isolation by 80 people), and was not released with an open license. In Tatman and Kasten (2017), the authors evaluate Youtube captioning and Bing Speech API on the Dialects of English Archive (Meier, 2019).

Even though these works are valuable in their findings, neither of the datasets used are released under a Creative Commons license, which greatly limits their usefulness. Furthermore, these corpora are not representative of the kind of speech seen in consumer-facing speech technology. The Artie Bias Corpus addresses these issues by releasing crowd-sourced speech data under an open license.

The techniques and tools outlined here can be easily ported over to any language in Common Voice. English currently dominates speech technology research, but it is in many ways a typological outlier. Generalizations from English to all languages should be taken with a grain of salt (Bender, 2009). We hope future work will extend the Artie Bias Corpus to other languages.

3. Corpus Design

The Artie Bias Corpus consists of 1,712 audio clips (≈ 2.4 hours) along with their transcripts and (self-identified, opt-in) demographic data about the speaker. The Artie Bias Corpus is in most cases useful as a test set, rather than a training or validation set.

The validity of these transcripts was first vetted by crowd-sourced votes on the Mozilla Common Voice platform, and then a second-round of vetting was performed by trained experts (the authors of this paper). This two-step validation process results in a much higher certainty of the transcripts. The validity of the demographic tags was not able to be investigated.

3.1. Filtering Common Voice

The Artie Bias Corpus is a subset of the test set of the English Common Voice corpus, which was released on June

[1] The Artie Bias Corpus is available for download: `ml-corpora.artie.com/artie-bias-corpus.tar.gz`

[2] Code for performing bias detection is available here: `https://github.com/artie-inc/artie-bias-corpus`

12, 2019[3]. The Common Voice corpus was collected and validated via a web interface.[4] Collection was performed by presenting text sentences on the screen, and capturing audio from the user's microphone as they read the text. Validation was performed by listeners confirming or rejecting the validity of a <audio,transcript> pair. After two volunteers confirm that a clip matches its transcript, the pair is marked "valid". If a clip first reaches two down-votes, the recording is marked as "invalid" and is not included in the train, test, or development datasets. Participants were prompted (though not required) to give a small amount of demographic information: "accent", "age", and "gender".

The Artie Bias Corpus contains all clips from this test corpus which are labeled with at least one piece of demographic information. The resulting subset contained 1,903 English test utterances. These 1,903 sentences have already been validated by volunteers, but we found that these validated clips contain numerous false-positives.

3.2. Annotation of Artie Bias Corpus

We re-validated these 1,903 testset audio clips and transcriptions with a team of trained native speakers of American English[5] (i.e. the authors of this paper). Only clips which passed with two out of three expert votes were included in the Artie Bias Corpus.[6] We validated the audio clips with a fork of the original web interface.[7]

The Artie Bias Corpus contains a total of 1,712 audio clips (≈ 2.4 hours). As a result of validation, we removed 167 clips. The Kappa statistic for our inter-annotator agreement can be found in Table (1) for our top three pairs of annotators. As per Viera et al. (2005), we reached "fair" to "moderate" agreement. Given that all annotators were native speakers of American English, we looked at the summary statistics of our rejections to identify annotation bias towards a certain demographic, and these statistics can be found in Appendix (1). We do not find strong evidence for an annotation bias against a certain demographic.

Table 1: Inter-annotator Agreement: Kappa Statistic

Annotator Pair	Number of Shared Clips	Kappa Statistic
<A,B>	1449	0.51
<A,C>	1014	0.28
<B,C>	620	0.29

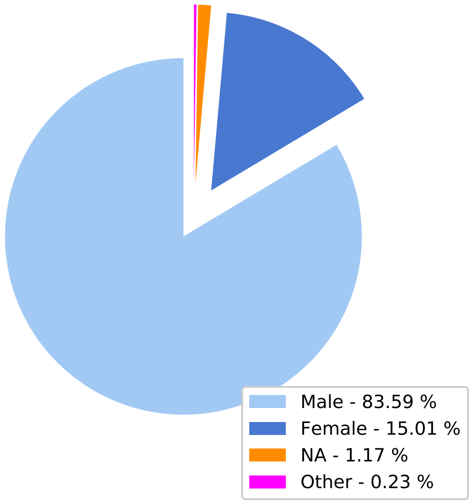
In addition, we excluded 22 clips containing children's speech, 1 clip which was sung, and 1 clip which contained problematic text. For privacy concerns we do not include children's speech in the Artie Bias Corpus.

3.3. Corpus Demographics

The Artie Bias Corpus contains information on 3 gender classes, 17 English accents, and 8 age ranges. For each demographic dimension, there is the possibility that participants chose not to share their information. As such, there is also the "NA" label present for many clips.

With regard to gender we find that the Artie Bias Corpus is heavily skewed towards male contributors (1,431 clips) over females (257 clips). We also observe 20 clips without gender information (i.e. "NA") and 4 clips whose speakers marked gender as "other". Figure (1) displays the proportions of these gender classes. With regard to the age of speakers, we find that the Artie Bias Corpus is skewed towards people in their twenties (c.f. Figure (2)).

Figure 1: Gender Distribution in the Artie Bias Corpus.



With regard to accents in the Artie Bias Corpus, the set of labels only contains English accents linked to country or geographic region (e.g. American English vs. Welsh English). Most of these labels are tied to countries in which English is an official language. There was no option for "non-native" accent, but our impression is that a large percentage of speakers had non-native accents, and these speakers may have marked their accent as "other" or left the field "NA". Identifying the accent of a speaker without knowledge of speaker identity is a non-trivial problem, and we chose to not make judgments on the validity of these accent tags.

As shown in Figure (3) and Appendix (1), the most common label for accent in the Artie Bias Corpus is "NA" (562 clips). In a close second comes speakers of United States English (558 clips). The third most common accent is Indian English (264 clips) and then English English[8] (131 clips), and all other accents have less than 50 clips each.

4. ASR Experiments

In order to test the capabilities of the Artie Bias Corpus in detecting bias in speech applications, we choose ASR as a

[3] This Common Voice release can be downloaded here: https://voice.mozilla.org/en/datasets

[4] The web interface can be accessed here: https://voice.mozilla.org/

[5] Ideally we would have a native speaker of each accent validating only audio from their accent, however, this is not feasible in the scope of the current project.

[6] Our annotation guide can be found in our code repository: https://github.com/artie-inc/artie-bias-corpus

[7] The code for the validation web app can be found here: https://github.com/artie-inc/voice-web

[8] "English English" is not necessarily United Kingdom English, given that Welsh English, Scottish English, and Irish English were all separate tags.

Figure 2: Age Distribution in the Artie Bias Corpus (not shown: "NA")

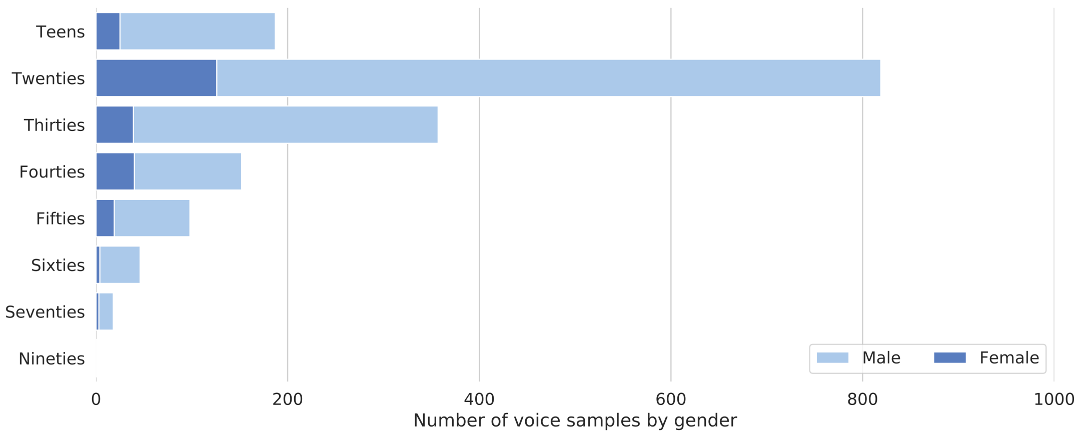
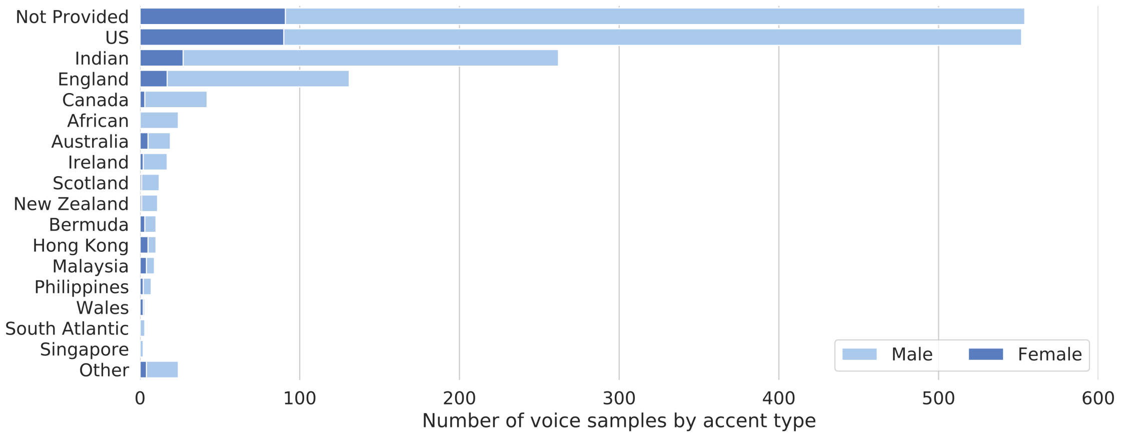


Figure 3: Accent Distribution in the Artie Bias Corpus



use case.[9] We demonstrate the ability of the Artie Bias Corpus to detect bias in the v0.5.1 release of Mozilla's DeepSpeech[10]. We then attempt to mitigate that bias via finetuning and detect that improvement with the Artie Bias Corpus.

For this study we investigate an open model for two reasons. Firstly, because using open models is good for replicability, and secondly because using closed models leads to dubious conclusions. Closed models are updated on an unknown schedule and any results have little lasting power. We have already seen conflicting results on gender bias in Youtube caption systems (Tatman, 2017; Tatman and Kasten, 2017), and there are too many unknowns to determine the cause of this discrepency. Furthermore, we don't have knowledge of the demographics of closed training corpora. Using free and open ASR models increases the replicability and accessibility of this research.

Our primary metric of ASR model quality is the Character Error Rate (CER)[11], computed as percentage of incorrectly transcribed characters as per the Levenshtein distance between the source transcript and the predicted transcript (Fiscus et al., 2006). For a sufficiently large corpus, the CER of individual samples will follow a normal distribution $\mathcal{N}(\mu, \sigma)$. The parameters μ and σ depend on both the model and corpus.

Given a partition of a corpus into C_1 and C_2 (for example, by demographic), we define bias in an ASR model between C_1 and C_2 as the difference in the two distributions $\mathcal{N}(\mu_1, \sigma_1)$ and $\mathcal{N}(\mu_2, \sigma_2)$. Given a partitioned corpus $C = C_1 \cup C_2$ and samples S_1 and S_2 from the two corpora, we can used the well-studied ANOVA null-hypothesis test to determine whether two samples have a significantly different CER.

[9] The code needed to perform statistics can be found here: `https://github.com/artie-inc/artie-bias-corpus`

[10] DeepSpeech v0.5.1: `https://github.com/mozilla/DeepSpeech/releases/tag/v0.5.1`

[11] We report Character Error Rate as opposed to Word Error Rate because the former is more language-agnostic (i.e. useful for non-English languages). Many languages do not use whitespace to delimit words, and there is no clear definition of "word" in a multilingual context. Even though we present data for English, we encourage our methods to be applied to more languages for which words are not an appropriate level of measurement.

4.1. Background on Baseline Model

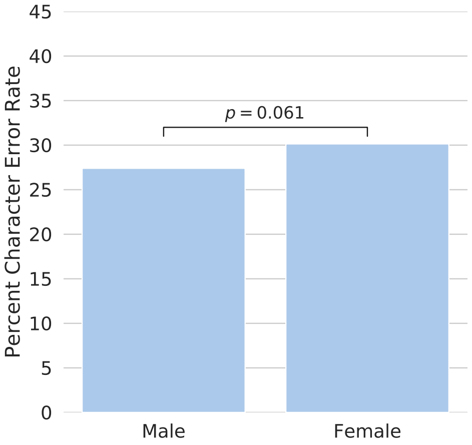
Our baseline model is the off-the-shelf v0.5.1 release of Mozilla's DeepSpeech. This model architecture was introduced by Baidu (Amodei et al., 2016) and implemented by Mozilla (Morais, 2018). This DeepSpeech is a 6-layer neural network with a single uni-directional LSTM layer, trained end-to-end via the Connectionist Temporal Classification (Graves et al., 2006) loss function.

This baseline model was trained on three large corpora: Fisher (Cieri et al., 2004), LibriSpeech (Panayotov et al., 2015), and Switchboard (Godfrey et al., 1992). The Fisher Corpus is heavily skewed towards North American English, and slightly skewed towards women (53% female). There are no statistics on the accents of LibriSpeech, but it was explicitly designed to skew towards US English, and is slightly biased towards men (48% female). The Switchboard corpus is entirely US English, and also slightly biased towards men (44% female). Given that DeepSpeech was trained on these corpora, we expect it to be biased towards US English, but it's hard to make an *a priori* prediction on gender bias.

4.2. Bias Detection Experiments

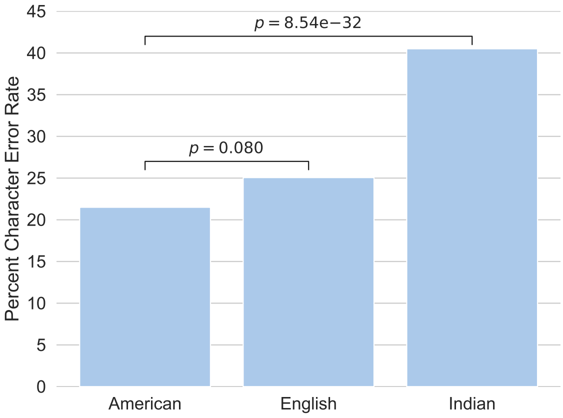
To investigate the bias in our baseline, we choose to limit our statistical analyses to five major demographic groups: women, men, Indian accents, English accents, and American accents. These are the groups for which we have most data in the Artie Bias Corpus. Given the training data used in the baseline model, we have reason to believe an accent bias exists. Given previous research on gender bias in ASR, we are interested in whether or not gender bias exists in this particular baseline.

Figure 4: Gender Bias in Baseline Model: We do not find evidence for a statistically significant difference between error rates for male vs. female speakers. Results come from testing on the Artie Bias Corpus.



We decoded all of the Artie Bias Corpus with the aforementioned off-the-shelf release of DeepSpeech (using the pre-trained n-gram language model). Results are shown for gender in Figure (4) and for accent in Figure (5). We do not find a statistically significant difference between men and women ($p = 0.061$), with women having a mean CER of 30.14% and men having a mean CER of 27.40%. We do, however, find a significant difference ($p = 8.54\text{e}{-32}$) between Indian English (40.50% CER) and US English (21.50% CER). We did not find evidence for a statistically significant difference ($p = 0.080$) between US accents and English accents (25.06% CER).

Figure 5: Accent Bias in Baseline Model: We find that the baseline model performs significantly better on American English speakers compared to Indian English speakers. However, we do not find a significant difference in performance on speech from England vs. the United States. Results come from testing on the Artie Bias Corpus.



In summary, we found that this baseline model does exhibit bias with regard to accent: better performance on US English compared to Indian English. However, we did not find evidence for a gender bias (men vs. women), or a bias for US English vs. English English. In the following experiments, we attempt to mitigate this found accent bias via fine-tuning, and measure improvement with the Artie Bias Corpus. We also demonstrate that it is possible to inadvertently create bias via fine-tuning.

4.3. Bias Mitigation Experiments

In the following experiments we attempt to mitigate the bias detected in the previous section. We demonstrate that fine-tuning on a target demographic can be an effective method in bias mitigation. All results come from testing on the Artie Bias Corpus.

4.3.1. Fine-tuning Datasets

We create the fine-tuning datasets shown in Table (3) from the English Common Voice train and development sets. We did not perform any extra validation of these <audio,transcript> pairs. The "All CV" row refers to all audio from Common Voice, including clips which have no demographic labels. The rows following (i.e. "male", "female", "American", "Indian", "English") only include audio clips from the designated demographic group.

The fine-tuning train and development sets are not included in the Artie Bias Corpus. The Artie Bias Corpus refers exclusively to the test set.

Table 2: Bias Mitigation Results (% Character Error Rate): Each column represents a subset of the Artie Bias Corpus, and each row represents a different DeepSpeech model. A **bolded** value represents a fine-tuned model which performed significantly better than the baseline on a certain demographic ($p < 0.05$ as per one-way ANOVA). The baseline is an off-the-shelf release, and the following rows were fine-tuned as described in Section (4.3). We observe improvement on all demographic groups except for English English. Blank cells represent values which are not of interest in the current study (i.e. demographic interaction effects).

		Testing Condition				
		Male	Female	American	Indian	English
Baseline	Mozilla Release	27.40	30.14	21.50	40.50	25.06
Initial Fine-Tuning	+ All CV	**23.70**	27.25	19.96	**34.45**	22.05
Secondary Fine-Tuning	+ Male	**22.69**	**25.80**			
	+ Female	**23.09**	**25.91**			
	+ American			**18.90**	**33.48**	21.26
	+ Indian			19.44	**32.77**	21.65
	+ English			19.33	**33.90**	21.07

Table 3: Number of audio clips per Training and Development set.

	Fine-Tuning Subset	
Demographic	**Train**	**Dev**
All CV	62,718	8,013
Male	12,957	4,364
Female	9,858	766
American	24,788	1,931
Indian	4,527	501
English	5,238	447

4.3.2. Experimental Method

The following experiments come from a two-step fine-tuning approach. First, we fine-tune the v0.5.1 pre-trained release of Mozilla's DeepSpeech to all of Common Voice. This first-pass of fine-tuning was performed to adjust the entire model to the general characteristics of the Artie Bias Corpus (e.g. speaking style, noisiness), which are significantly different from the training corpora used for the pre-trained model. We use a dropout rate of 15% and a learning rate of $1\mathrm{e}{-5}$ over 100 epochs. The dropout rate was chosen to match that of the original DeepSpeech training, and the learning rate was found to be best over a sweep from $1\mathrm{e}{-3}$ to $1\mathrm{e}{-6}$. Second, we perform further fine-tuning to each target demographic individually, with an extra 20 epochs of backpropagation with a dropout rate of 15% and a learning-rate of $1\mathrm{e}{-6}$. We end up with a total of seven models as a result: 1 baseline model, 1 model fine-tuned to Common Voice, and 5 models fine-tuned to a certain demographic (male, female, US accent, Indian accent, English accent).

4.3.3. Results

Results from the seven ASR models are displayed in Table (2). The rows represent different models and the columns represent demographics from the Artie Bias Corpus. In this section we are interested in differences between models rather than intra-model differences. Intra-model differences (i.e. bias) are investigated in Section (4.2).

In the first row of Table (2) we report results from the baseline Mozilla DeepSpeech model. This is the same model in which we identified bias in Section (4.2). The following experiments are an attempt to decrease the percent CER on each demographic via fine-tuning.

In the second row of Table (2) we present our results after fine-tuning the baseline to the entire Common Voice corpus (hence the row label "All CV"). This model was able to significantly outperform the baseline on male speakers and Indian English speakers, but there was no significant difference for women, US English, or English English.

Lastly, in rows 3 through 6 of Table (2) we performed further fine-tuning to individual demographic subsets. For example, the row headed by "Female" represents a model that was initially fine-tuned to all audio, and then fine-tuned to only female speakers in Common Voice. With this extra step of fine-tuning we improve performance on women and US compared to the baseline. The improvements in male speech and Indian English still hold from the initial fine-tuning to all audio. We are never able to find significant improvement for English English.

4.3.4. Creating Unintended Bias

Even though we made improvements for individual demographic groups, in one case we inadvertently created significant bias. In the baseline model, we did not originally find evidence for a gender bias even with a 2.74% CER difference between male and female group means (c.f. Figure (4)). However, after fine-tuning this difference became significant.

After an initial fine-tuning to all of Common Voice, we find a significant difference in accuracy on gender ($p = 0.042$), with a gap in CER at 3.55%. This finding actually makes sense, given that we first fine-tune to a male-skewed dataset (i.e. all of Common Voice). After secondary fine-tuning to female voices the gender gap decreases to a 2.8% difference in CER, but the gap remains significant.

5. Limitations

There are two main limitations we have identified with regard to this study. Firstly, we have a statistical power limitation. As mentioned in Section (4), ANOVA statistical tests in this scenario are dependent on both the model and the data. As the size of the dataset tends to infinity, the test results are more likely to reflect true bias in the model. However, the smaller the dataset, the less power we have to detect a true bias. In order to detect a true bias with a small dataset, the bias should be large and reliable (i.e. small standard deviation within a group). For instance, it may be the case that the baseline model is biased, but that the Artie Bias Corpus is not large enough to detect that bias with statistical certainty.

A second limitation of the Artie Bias Corpus is the possibility of spurious correlated signals. Demographic correlation could exist such that one demographic dimension is correlated with another, resulting in misleading statistics. With regard to the two demographic dimensions investigated here (i.e. gender and accent), we do not find evidence for a correlation (c.f. Figure (3)). There are more possible interaction effects which should be investigated further. With regard to other factors, it may be the case that given regional internet and hardware there are specific kinds of noise correlated with demographic groups. This is a difficult case to disentangle, but may be worth investigating.

6. Concluding Remarks

In this paper we introduced the Artie Bias Corpus – a collection of<audio,transcript> pairs released under a Creative Commons license. The Artie Bias Corpus is intended for demographic bias detection in speech technologies such as Automatic Speech Recognition and Speaker Verification.

We outlined the design, annotation, and content of the corpus in Section (3). We demonstrated that the Artie Bias Corpus is capable of detecting demographic bias in Automatic Speech Recognition applications in Section (4.2). We further demonstrated the corpus is able to detect improvements in model performance on demographic groups in Section (4.3). This is the first version of the Artie Bias Corpus, and we welcome its growth to more languages and more data.

Acknowledgments

We would like to thank Dr. Natasha Warner for a very useful discussion on statistical hypothesis testing. We also thank Michael Henretty and Gregor Weber for help with setting up the `voice-web` validation web app. Any shortcomings remain our own.

7. References

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182.

Bender, E. M. (2009). Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece, March. Association for Computational Linguistics.

Cieri, C., Miller, D., and Walker, K. (2004). The fisher corpus: a resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

Fiscus, J. G., Ajot, J., Radde, N., and Laprun, C. (2006). Multiple dimension levenshtein edit distance calculations for evaluating automatic speech recognition systems during simultaneous speech. In *LREC*, pages 803–808. Citeseer.

Garnerin, M., Rossato, S., and Besacier, L. (2019). Gender representation in french broadcast corpora and its impact on asr performance. In *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery*, pages 3–9. ACM.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520. IEEE.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 369–376, New York, NY, USA. ACM.

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. (2018). Fairness without demographics in repeated loss minimization. *arXiv preprint arXiv:1806.08010*.

Meier, P. (2019). International dialects of english archive. `https://www.dialectsarchive.com`.

Morais, R. (2018). Streaming rnns in tensorflow, Sep.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Tatman, R. and Kasten, C. (2017). Effects of talker dialect, gender and race on accuracy of Bing speech and YouTube automatic captions. In *Proceedings of Interspeech 2017*, pages 934–938.

Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. *Ethics in Natural Language Processing*.

Viera, A. J., Garrett, J. M., et al. (2005). Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363.

A Demographics of Removed Clips

The data found in Table (4) display the summary statistics on the amount of audio clips removed from the Common Voice corpus to create the Artie Bias Corpus. We find higher than average rejection rates for clips marked with "Other" for either accent or gender, but we do not find a large difference between men and women rejections, and we do not find a large difference in rejection rates between the three largest accent groups: US English, Indian English and English English.

Table 4: Statistics on Removed Clips per Demographic

Demographic	Original Number of Clips	Number of clips in Artie Bias Corpus	Percent Clips Removed
Teens	221	187	15.38
Twenties	900	827	8.11
Thirties	410	366	10.73
Fourties	171	152	11.11
Fifties	107	101	5.61
Sixties	53	46	13.21
Seventies	20	18	10.00
Eighties	3	0	100.00
Nineties	1	1	0.00
NA	17	14	17.65
Female	281	257	8.54
Male	1591	1431	10.06
Other	9	4	55.56
NA	22	20	9.09
African	25	24	4.00
Australia	20	19	5.00
Bermuda	10	10	0.00
Canada	55	42	23.64
England	151	131	13.25
Hongkong	11	10	9.09
Indian	281	264	6.05
Ireland	23	21	8.70
Malaysia	11	9	18.18
New Zealand	11	11	0.00
Philippines	10	7	30.00
Scotland	12	12	0.00
Singapore	2	2	0.00
South Atlantic	3	3	0.00
United States	616	558	9.42
Wales	3	3	0.00
Other	33	24	27.27
NA	626	562	10.22