Responsible Benchmarking of Fairness for Automatic Speech Recognition

Felix Herron[(1,2)], **Ange Richard**[‡(2,3)],
François Portet[‡, (2)], **Alexandre Allauzen**[(1)], **Solange Rossato**[‡, (2)]
MILES Team, LAMSADE, Université Paris Dauphine-PSL (1)
GETALP Team, LIG, Université Grenoble Alpes (2)
PACTE, Université Grenoble-Alpes (3)
felix.herron@univ-grenoble-alpes.fr

Abstract

Many studies have shown automatic speech processing (ASR) systems have unequal performance across speaker groups (SG's). However, the manner in which such studies arrive at this conclusion is inconsistent. To pave the way for more reliable results in future studies, we lay out best practices for benchmarking ASR fairness based on literature from machine learning fairness, social sciences, and speech science. We then perform a case study on the Fair-speech benchmark, applying aforementioned best practices, and discuss how failing to do so can result in erroneous conclusions. On the whole, we advocate for as fine-grained an analysis as possible, taking into account as many variables as are available, in order to eschew dataset-level bias.

Keywords: Fairness, benchmarking, statistics

1. Introduction

In recent years, automatic speech recognition (ASR) software has grown increasingly performant (Nayeem et al., 2025), which has led to a complementary increase in prevalence of ASR use among diverse populations (Yang et al., 2024; Wald et al., 2024; Dino et al., 2025). It is therefore increasingly imperative to ensure that existing ASR systems perform equally regardless of the identity of the speaker. There is ample research demonstrating that certain speaker groups (SG's), such as children and non-native speakers, are treated worse than others by ASR systems. However, the methodology for identifying such SG bias in ASR systems is inconsistent across studies and sometimes marred by lack of precise analysis of the multifaceted identities of individual speakers.

Indeed, applying fairness benchmarks without sufficient oversight can lead researchers to stumble into erroneous conclusions which are incongruous with real world biases, a common blunder in fairness research (Selbst et al., 2019). This paper discusses the importance of defining SG-level fairness as intentionally and precisely as possible to avoid such blunders. We start by diagnosing our observed lack of consistency across ASR fairness benchmarking studies, and suggest this is due in part to a lack of clarity about how SG-level fairness should be defined and measured. We then list several best practices to avoid accidentally measuring bias stemming from fairness corpora rather than real-world bias. We formally define fairness in ASR, as motivated by broader fairness literature in machine learning (ML). Finally, we perform a case study on a common fairness corpus, Fair-speech, and apply many of the best practices to our analysis. We highlight some pitfalls that could entrap unaware users of the corpus. We finish by suggesting how future work, particularly in dataset creation, can help facilitate fairness benchmarking in ASR.

‡ contributed towards formulating framework of speaker group intersectionality and multivariate SG's.

2. Motivation

This study is motivated by a lack of consistency both within single and among studies examining fairness in SOTA ASR systems. We were puzzled to find that some studies report that women experience *significantly worse* treatment (Hutiri and Ding, 2022; Garnerin et al., 2021; Tatman and Kasten, 2017; ElGhazaly et al., 2025), *significantly better* treatment (Veliche et al., 2024; Feng et al., 2024, 2021; Abushariah and Sawalha, 2013), or *both*, depending on subgroup studied (Attanasio et al., 2024; Tatman, 2017; Kulkarni et al., 2024). Likewise, while most studies find that non-native speakers are *less well understood* by ASR systems (Ghorbani and Hansen, 2018; Zhang et al., 2022; Sekkat et al., 2024), some *find the opposite* (Veliche et al., 2024). Studies seem to all agree that *children receive worse performance* by ASR systems; however, whether older adults are better understood than younger *varies broadly by publication* (Aman et al., 2013; Feng et al., 2021; Kulkarni et al., 2024; Sekkat et al., 2024).

It is possible that this effect is due in some part to

Proceedings of SPEAKABLE @ LREC 2026, pages 66–78
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

the different ASR models being evaluated by each of these studies. Fixing a dataset and two SG's g_1,g_2 (e.g. native vs non-native speakers), some studies find that different ASR models perform better on g_1 while others on g_2 (Attanasio et al., 2024). However, most studies in the literature find that different ASR systems tend to be biased against the same SG's (Feng et al., 2024, 2021; Fuckner et al., 2023). If we assume that all ASR systems reflect the same SG-level biases, we would hope that all studies into ASR fairness would arrive at similar conclusions. That they don't is therefore likely due to methodological variance which allows researchers analyzing the same data to arrive at different conclusions.

3. Best practices in reducing transmission of dataset bias

It is important to remember that any conclusions reached by fairness studies are *estimates* of real life bias, influenced by data on which the experiments were performed. With this in mind, researchers should try and limit dataset-level bias as much as possible so that their estimates are as close as possible to a real world simulation. In this section, we will highlight several key notions to attenuate filtration of biases due to dataset construction into fairness results. For each, we cite examples from the literature. It is important to note that these best practices remain general - we see them as akin to tools in a belt, whereby the user still must know how to use each in the manner most beneficial to their use case. The subsequent section describes a case study on how these tools can be put to use - however, each setting is different.

3.1. Ensure equal distribution of recording quality

Background noise (and recording quality in general) have been shown to impact ASR performance (Rodrigues et al., 2019). It is possible that some SG's will have different levels of background noise, thus potentially biasing results of a fairness study. Some benchmarks circumvent this by recording all of their inputs in the same conditions (Sekkat et al., 2024; Veliche et al., 2024), though benchmarks compiled from diverse sources cannot (Meyer et al., 2020). That said, it can be useful to have variably noisy recordings in the dataset, as real life ASR often occurs in noisy environments using old recording equipment. Furthermore, some SG's are more likely experience such potential impediments to ASR transcription, such as the "low" socioeconomic status SG in the Fair-speech dataset (Veliche et al., 2024).

3.2. Verify text complexity

If different SG's use different vocabulary/grammar (i.e. text) in the real world, it is important that these attributes be taken into account during bias testing. A corpus where all speakers speak comparable texts cannot be considered to faithfully estimate bias if this is not the case in the real world; likewise, if a certain SG in a corpus is comprised of recordings of complex texts, while another SG speaks easy texts, this likewise has the potential to engender unfaithful bias estimations. On the other hand, if one is trying to capture *merely* the bias due to acoustic features of a speaker's voice, then controlling for text (complexity) is beneficial. Once again, this is a decision that researchers must consider intentionally in the context of their individual study.

For example, Koenecke et al. (2020) proposes calculating each text's perplexity to assess whether each SG speaks similarly complex sentences. They also measure the "dialect density" of text to determine the extent to which it contains grammar typical of African American Vernacular English.

3.3. Understand intra-SG speaker diversity

When we measure ASR performance w.r.t. a SG, it is tempting to treat SG labels as precise, immutable defining characteristics of speakers. However, speakers within a given SG can be diverse. It is therefore essential to understand how each SG defined, what it means to belong to multiple SG's at once, and SG's are balanced throughout the dataset.

3.3.1 Intersectionality Traditionally, fairness in ML has focused on comparing between outcomes for single groups from the same demographic variables (DV's), such as between different races or sexes (Bellamy et al., 2018; Friedler et al., 2018). However, recently the application of **intersectionality** in ML fairness research has gained traction as a technique to more precisely gauge bias (Foulds et al., 2019; Wang et al., 2022). These studies argue that measuring fairness w.r.t. a single DV in isolation is insufficient; to best understand fairness, one must look at as fine-grained SG's as possible, comprised of as many DV's as possible.

Intersectionality has its roots in the social sciences where Crenshaw (1989) defines it as discrimination faced by members of multiple marginalized classes at the *intersection* of several groups, for example Black. Wang et al. (2022) emphasizes that treating heterogeneous groups as a monolith (e.g. all people from Pacific islands) can hide unfair treatment experienced by subgroups thereof (e.g. specific islands). Foulds et al. (2019) emphasizes the importance of prioritizing protected classes which are underrepresented in fairness

benchmarks, as their discrimination can be more easily ignored than a large underprivileged group.

Several existing benchmarks for fairness in ASR already consider the intersectionality of SG's (without necessarily using that exact term) for both aforementioned motivations. For example, Feng et al. (2021) and Feng et al. (2024) examine the WER gap between Dutch spoken in Flanders vs in the Netherlands, intersecting with regional dialects, native-ness of speakers, age, and speech format (read vs. human-machine interaction). One finding is that the gender-based WER gap is least significant among children, while the age-based WER gap is most significant for women, as well as that the regional gap is strongest among children and teens, and weakest among older adults. This is a crucial insight that would be ignored if the authors had only compared between genders, ages, or regions.

3.3.2 Conditional statistical parity The metrics used to measure fairness in ASR correspond to **statistical parity** as introduced by Verma and Rubin (2018) in their landmark fairness taxonomy paper. A more rigorous version of this is **conditional statistical parity**, which requires that all secondary attributes about the setting be the same, such as background noise or text complexity. Furthermore, we can condition on/take into account secondary DV's in our calculations. This is important both in order to uncover intersectional biases, as well as to equilibrate potential unbalance in other DV's. Failure to do this might end up spuriously measuring the random side-effects of subgroup imbalance captured during dataset construction.

As a toy example, let us imagine we are measuring the fairness of performance of an ASR system on men vs women, on some benchmark B. By chance, 1 in 20 men in B suffer from Parkinson's disease (diminishing their ASR comprehensibility (Moro-Velázquez et al., 2019)), but only 1 in 100 women in B suffer from Parkinson's. If we are either unaware of this, or ignore it, then we might conclude that men experience worse ASR performance than women; however, our observation would at least in part be due to the confounding influence of a higher prevalence of Parkinson's in the male population, rather than due to their masculinity.

Koenecke et al. (2020) responsibly attempt to avoid contamination of their race DV by either age and gender by retaining the same proportion of both gender and age groups for both White and Black speakers. However, they don't consider the intersection of age and gender, an oversight of intersectionality. They then delve into the geography of both racial groups where, crucially, they find that race is not sufficient to explain the WER gap; Black Americans from Rochester had comparable WER to White Americans.

Sekkat et al. (2024) control for the confounding effects of other DV's, noting that this causes some univariate effects w.r.t. age and gender to vanish. Likewise, Tatman (2017) finds a greater difference in by-gender performance in certain dialect groups.

On the other hand, Veliche et al. (2024) finds that men have twice as high a WER as women, which they explain by citing previous work showing men tend to have worse ASR performance. However, while they note that the men in their dataset are far more likely to be African-American than their women, they don't perform an intersectional analysis to interrogate what is likely at least partially responsible for that effect.

Another example of failure to take SG diversity into account is the "Asian" dialect category in the Sonos dataset (Sekkat et al., 2024), which is comprised of speakers from Southern as well as Eastern Asian countries. The authors acknowledge the extraordinary diversity of this category and the incumbent challenges this causes in interpreting results based on it.

3.3.3 Beware (un)known confounding factors In the previous examples, researchers were able (or failed) to avoid jumping to conclusions more reflective of biases in their dataset than biases in their ASR systems. Or they were able to uncover intersectional bias specific to multidimensional SG's. However, these effects can only be explicitly controlled when potentially confounding metadata are available in the dataset. For example, the Sonos dataset has ethnicity tags for only a small portion of its speakers, where they show it to have a statistically significant relationship with ASR error rate (Sekkat et al., 2024). However, they cannot control for equal ethnicity distribution over the rest of the dataset, and therefore cannot control this bias.

Indeed, there may be many other confounding variables related to speaker identity which are not included in the dataset. It is by definition impossible to directly control for these; however, authors could estimate the extent to which their datasets are free from such effects by using phonetic priors. For example, studies have shown that human's voices don't change very much during middle age (Rojas et al., 2020). The reliability of fairness experiments can therefore be benchmarked by performance variance across middle-aged age groups: if there is significant performance difference between any two middle-aged SG's, that is an indication of methodological error, likely due to lack of balancing (Sekkat et al., 2024; Veliche et al., 2024).

3.4. Define SG-level performance based on speaker-level performance

When measuring SG-level bias, it is imperative to calculate error for each SG as a function of error for

each speaker adhering to said SG. This is based on two observations: first, utterances from the same speaker are not independent, and thus we cannot perform a statistical test that assumes independence of samples. Second, this avoids bias due to imbalance in representation for each speaker. For example, if speaking time is not equally distributed across speakers in SG (e.g. D'_{SG} contains many more utterances/words for some speaker S_1 than another speaker S_2), then calculating SG-level error as a function utterances will engender bias towards Speaker S_1, and will not be a faithful representation of the SG overall.

Not all studies adhere to this principle, however. Sekkat et al. (2024) explicitly argues for the simplicity of measuring fairness based on individual utterances. Furthermore, Feng et al. (2021) and Feng et al. (2024) base some of their conclusions on a small numbers of speakers (see i.e. Table 1 in Feng et al. (2024)). They claim statistical significance, likely based on the number of overall samples or hours of recorded speech, rather than the diversity of speakers per SG.

3.4.1 The challenge of speaker paucity If we define SG-level performance as a function of individual speakers, the statistical significance of our results will depend on the number of speakers in each SG. This leads us to a set of contradictory incentives: the more precisely we define SG's as the intersections of multiple DV's (as encouraged in the previous section), the more precise are our conclusions into SG-level fairness. However, the more precisely we define SG's as the intersections of multiple DV's, the fewer speakers will be included in each class, thus reducing the significance of tests we perform on them. This is logical: if a SG contains too few speakers, we risk measuring bias due to the unique nature of those several individuals, rather than due to the SG they belong to.

We can derive the number n of speakers per SG necessary for statistical (with confidence α and power β) significance in, for example, a one-sided two-sample Z-test (e.g. comparing the mean error of two SG's given fixed population variance) with test-statistic Z:

$$
\begin{aligned}
Z &:= \frac{\hat{\delta}}{\sigma\sqrt{2/n}} \\
Z > z_\alpha + z_\beta &\implies \frac{\hat{\delta}^2}{\sigma^2 * (2/n)} > (z_\alpha + z_\beta)^2 \\
&\implies n > 2 * \frac{(z_\alpha + z_\beta)^2 \cdot \sigma^2}{\hat{\delta}^2} \qquad (1)
\end{aligned}
$$

where $\hat{\delta}$ is the difference in estimated WER for two SG's, σ the variance between speakers, and $z_{\alpha,\beta}$ the quantiles defined by the significance and power respectively. For example, given typical values of 95% confidence with 0.8 power, taking $\hat{\delta} = 0.1$ and $\sigma = 0.15$ (reasonable estimates based on our analyses in Section 5), we would need $n \approx 35$ speakers per SG. This could be a serious hindrance for corpora with few speakers, and/or multivariate SG's defined by many DV's.

This bound cannot be shrunk simply by increasing the number of utterances per speaker. $\hat{\sigma} := \sigma + \epsilon$, where $\epsilon > 0$ varies inversely to the number of words for each speaker - the more words available per speaker, the smaller ϵ and thus the less noisy $\hat{\sigma}$, which results in smaller n. However, $\hat{\sigma}$ can never be lower than σ, which means that an increase in the number of words per speaker has a limited effect on improving our measured error bounds.

3.5. (Sometimes,) aggregate SG's

Just because metadata are available in a corpus doesn't mean they are useful in fairness analysis!

3.5.1 Too few speakers per SG In this case, we will be unable to draw statistically significant conclusions based on performance over this SG. Therefore, it could make sense to create an "other" category which groups semantically unusual SG's together. The upside of such aggregation is that it allows "other" to be represented by sufficient speakers so as to be statistically significant, whereas as initially constituted, those SG's would be statistically meaningless. The downside is that this marginalizes the individual identities of those underrepresented SG's, which Wang et al. (2022) warns against. However, we are limited by the data available, and there is interpretable value in creating a class to compare with the mode SG's.

3.5.2 Superfluous level of precision in metadata We stand to gain nothing by measuring fairness w.r.t subgroups defined by attributes likely independent of ASR fairness, for example different middle-age subgroups (Rojas et al., 2020; Hustad et al., 2021). Instead, this can lead to two harms: 1) we risk accidentally creating SG's which are unbalanced w.r.t. other underlying SG's, thereby potentially corrupting our analysis. 2) it reduces the number of speakers in each multivariate SG, thereby lowering the statistical significance of results. Thus, we might consider aggregating all middle-aged speakers into the same SG.

3.6. Outlier speaker removal

A final source of bias potentially contaminating our SG-level fairness results are outlier speakers. If we want to want to measure the general behavior of a certain SG, which potentially contains few speakers (due to dataset constraints), and one of the speakers is understood much worse than the rest, that is potentially due to some individual speaker-level characteristics which are not relevant to our

analysis. It can therefore make sense to exclude extrema values from SG-level mean calculation, for example. This is less of a problem for datasets with very high numbers of speakers; however, that is unfortunately rarely the case in practice.

4. Quantifying fairness in ASR

We describe the two metrics most often used in the literature to quantify bias in ASR systems.

4.1. Relative SG-level error/WER gap

Most studies into SG-level ASR fairness measure the relative error rate for each SG. For a dataset D, they calculate the word error rate (WER) for each utterance, a measurement of the number of substitutions, deletions, and insertions necessary to correct the automatic transcription by an ASR model M over some $D' \subseteq D$. Some studies then calculate the average WER for each SG as the average WER for all utterances $D'_{SG} \subseteq D'$:

$$\text{WER avg.}^*(D'_{SG}) := M^*(D'_{SG}) := \frac{1}{|\{u \in D'_{SG}\}|} \sum_{u \in D'_{SG}} \text{WER}(u;M) \quad (2)$$

However, as mentioned in Section 3.4, Metric 2 falls immediately into a hazard of imprecise measurement by failing to first average by speaker. A more prudent approach is:

$$\text{WER avg.}(D'_{SG}) := M(D'_{SG}) := \frac{1}{|\{S \in SG\}|} \sum_{S \in SG} \frac{1}{|\{u \in D'_S\}|} \sum_{u \in D'_S} \text{WER}(u;M) \quad (3)$$

Then, one can measure bias against a particular SG SG_i in terms of the relative performance between SG_i and some the rest of the subset $D' \subseteq D$ (typically, studies use $D' = D$) (Veliche et al., 2024; Feng et al., 2024):

$$\text{Err}_{rel}(SG;D',M) := 100 \times \frac{M(D'_{SG}) - M(D')}{M(D')} \quad (4)$$

Some studies also calculate the unfairness w.r.t. a DV as the difference between the best and worst relative error for two constituent SG's, often denoted the WER gap (ElGhazaly et al., 2025; Attanasio et al., 2024; Kim et al., 2025):

$$\text{Unfairness} := \text{WER gap}(DV;D',M) := \max_{SG_i \in DV} \{\text{Err}_{rel}(SG_i;D',M)\} - \min_{SG_j \in DV} \{\text{Err}_{rel}(SG_j;D',M)\} \quad (5)$$

Then, one can perform a statistical test, such as a 1-sample (or 2-sample, if comparing between pairs of SG's rather than with respect to the dataset on average) t-test, to determine whether the relative error and unfairness are statistically significant for SG's and DV's respectively. An ASR model is deemed fair w.r.t. a SG if it delivers statistically negligible relative error, and fair w.r.t. a DV if it delivers a statistically negligible WER gap.

4.1.1 Isolated effect of single DV's As mentioned in Section 2, we recognize that individual SG's potentially contain heterogeneous subgroups. Thus we seek to isolate the effect of individual DV's on ASR performance. For any $DV^i \in [DV^1,..,DV^k]$ (e.g. gender), one of many DV's included in metadata, we define a subset $D'_{cond,i} \subseteq D$ by fixing a SG_{DV_j} for every other $DV_j \in [DV^1,..,DV^k] \backslash \{DV_i\}$ (e.g. only children, only non-native speakers, only African Americans, etc.):

$$D'_{cond,i} = \bigcap_{j \in [1..k] \backslash \{i\}} D_{SG_{DVj}} \quad (6)$$

and calculate $\text{Err}_{rel}(SG;D'_{cond,i},M)$ for all $SG \in DV^i$ as well as WER gap$(DV^j;D'_{cond,i},M)$. We can then aggregate mean performances for each permutation of other the other DV^j and perform a statistical test, such as 1-sample t-test, to determine, w.r.t. a $SG \in DV^i$, whether the means of the relative error rates were statistically significantly different from zero, or w.r.t. DV^i overall, whether the unfairness levels were statistically significantly greater than zero.

4.1.2 Worst treated multivariate SG's We define *multivariate SG's* as the intersection of every DV available in the dataset. For example, if the dataset is annotated with gender, age, and native language, an multivariate SG might be "female children who speak native English". For each multivariate SG, we calculate the relative error w.r.t. the dataset overall (Eq. 4); then observe which multivariate SG's, if any, have the lowest/highest relative performance. This will uncover SG's whose marginalization is compounded by their intersectionality, as proposed by Wang et al. (2022).

5. Case study on Fair-speech

Unfortunately, many of the most prominent corpora for evaluating ASR systems do not permit bias evaluation due to a lack of sufficient recorded demographic metadata (Ma et al., 2024; Panayotov et al., 2015; Linguistic Data Consortium, 2013) or unreliable labeling thereof (Ardila et al., 2020; Wang et al., 2024). However, there are several corpora specifically designed for bias/fairness evaluation of ASR systems whose multitudinous metadata categories permit finer-grained ASR evaluation. We

proceed to analyze the Fair-speech corpus (Veliche et al., 2024) and discuss how to implement some of the best practices from the previous section.

We replicate each experiment using three different near-SOTA ASR models: Whisper-medium (`Whisper`), wav2vec2-large-960h-lv60 (`Wav2vec 2.0`), and wav2vec2-large-xlsr-53-english (`XLS-R-En`). `Whisper` was trained end-to-end for ASR on 680k hours of YouTube transcripts (Radford et al., 2022); `Wav2vec 2.0` was pretrained on 60k hours of LibriLight (Baevski et al., 2020); `XLS-R-En` was pretrained on a multilingual corpus comprising 53 languages (Babu et al., 2021). `Wav2vec 2.0` was finetuned on 960h of LibriSpeech (Panayotov et al., 2015); `XLS-R-En` finetuned on the English split of CommonVoice (Ardila et al., 2020). We can thus test whether our results are specific to one architecture/training set, or general across ASR systems.

5.1. Dataset description

The Fair-speech Dataset (Fair-speech) comprises 593 speakers over 56 hours (Veliche et al., 2024). Fair-speech is comprised of recordings of paid speakers speaking (not reading) smart speaker commands. Speakers self-report metadata including: gender, age, ethnicity, first language, and socioeconomic background. See Fig. 1.

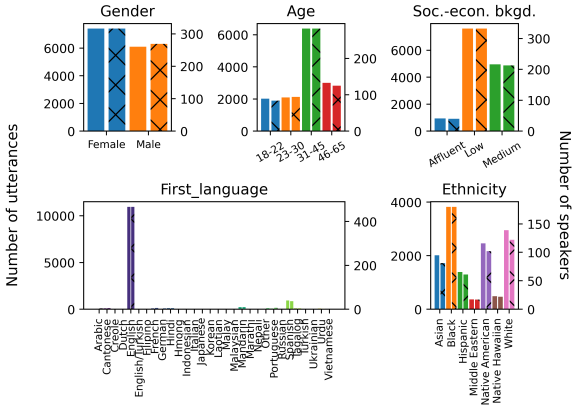

Fig. 1: Overall distribution of demographic labels in the Fair-speech corpus. Note that most speakers are native English speakers, with a small (illegible) minority of Spanish native speakers.

Fair-Speech represents many first languages; however, given a limited number of speakers per language, that might limit the statistical significance of results pertaining to speakers of sparsely represented languages. Furthermore, the number of age categories is likewise too precise - we likely stand to gain little by differentiating between different classes of middle-aged adults. Fair-speech lacks children and old adults, the two age-related SG's with phonetic motivation for divergent ASR error rates (Hustad et al., 2021; Rojas et al., 2020).

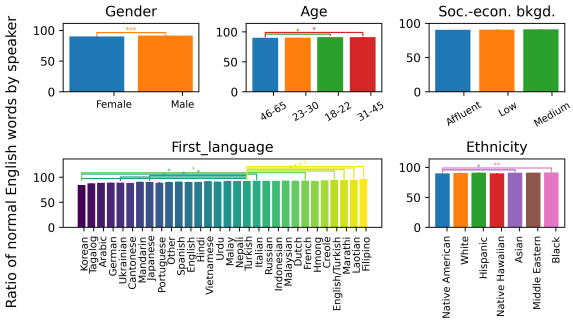

Fig. 2: Ratio of non-English words per sentence, averaged by speaker.

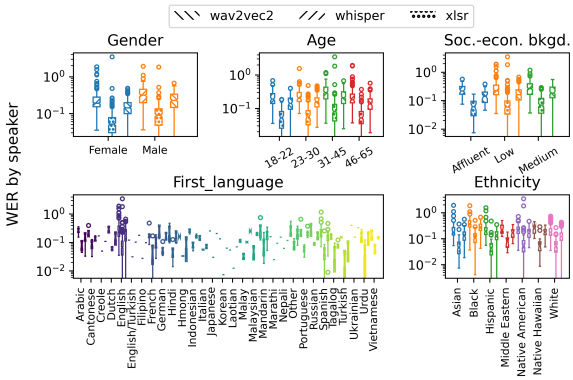

Fig. 3: Variance of WER for each ASR model, averaged by speaker.

5.2. Filtering out outlier speakers and utterances

First, we consider filtering out outlier speakers and utterances based on WER for each ASR model. Figure 3 shows the variance of WER for each univariate SG and model. Note that some speakers have an average WER of over 1 - this is likely due to speaker-specific anomalies and not an accurate reflection of the ASR system overall. We proceed by filtering out all speakers (and utterances per speaker) with a z-score of > 3 in each analysis that

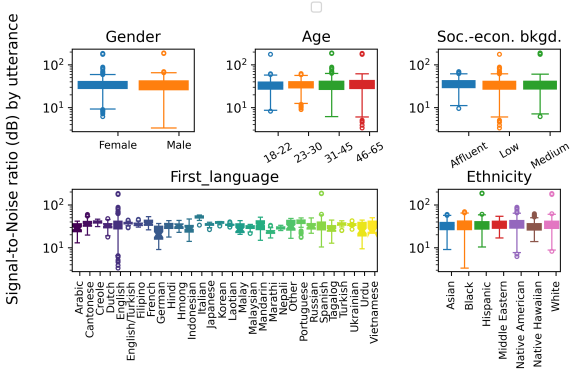

Fig. 4: Signal-to-noise ratio for each recording.

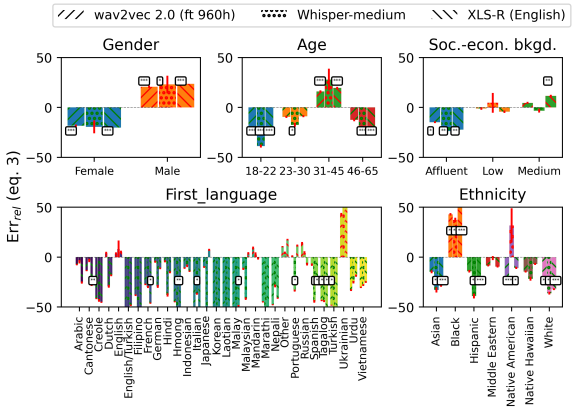

Fig. 5: Fair-speech when measuring relative WER gap between over SG's belonging to a single DV at once. Values < 0 indicate below average WER, i.e. above average performance. * denotes statistically significantly greater/less than 0 according to a 1-sample t-test - * implies $p < 0.05$, ** implies $p < 0.01$, *** implies $p < 0.001$.

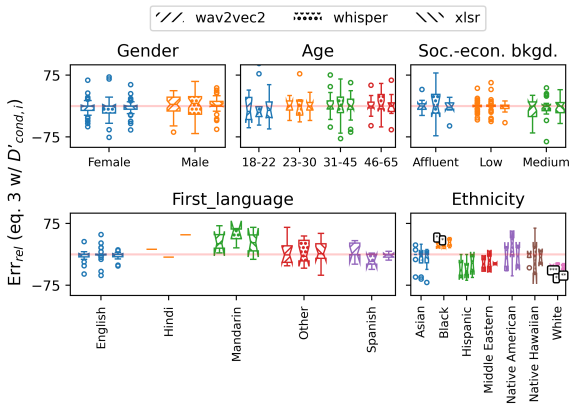

Fig. 6: Fair-speech when measuring relative WER gap between intersectional SG's differing only on one specific DV. Each datapoint is a statistically significant difference between SG's differing by only one DV. Values < 0 indicate below average WER (above average performance). * denotes statistically significantly greater/less than 0 in the aggregate according to a 1-sample t-test. * implies $p < 0.05$, ** implies $p < 0.01$, *** implies $p < 0.001$.

we conduct.

5.3. Recording quality and text complexity

Figure 2 shows the average text complexity for every speaker, measured by number of words in the transcript that are not standard English (we use NLTK English dictionary (Loper and Bird, 2002)). Overall, there is little variance, particularly among SG's with high representation; that said, the most

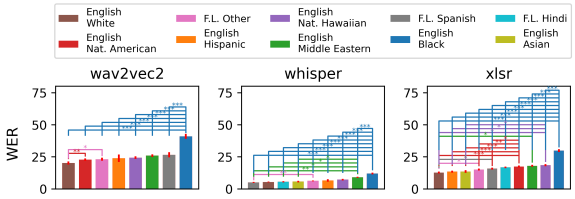

Fig. 7: Overall WER for Fair-speech when conditioning on first language and ethnicity. * implies denotes significantly higher WER on a one-sided, two-sample two-sample t-test (* implies $p < 0.05$, ** implies $p < 0.01$, *** implies $p < 0.001$).

disadvantaged SG's, as we will see in later analyses, are not necessarily those with the lowest ratio of standard vocabulary.

Figure 4 shows the signal-to-noise ratio of each utterance, broken down by SG. Note that most recordings have are 10 dB (a reasonable threshold for ASR performance (Bouchakour and Debyeche, 2022)). For our experiments, we will remove all recordings at $< 10dB$.

5.4. Calculating relative WER for each SG

We begin our case study by analyzing the results from Fair-speech. First, in Figure 5, we present relative error rate as the dataset was constructed, without conditioning or manipulation. We measure statistically significant performance discrepancies w.r.t. each of the five DV's recorded in Fair-speech. Our results correspond to what was initially published in (Veliche et al., 2024). Several odd results stand out, which raise some red flags about our experimental setup:

1. 31-45 year-old's have higher WER than all other age groups. This is likely evidence of poor subgroup balancing, as there is no logical reason for different age groups of middle-aged adults to have variant performance.

2. Men have vastly higher WER than women. Veliche et al. (2024) attempt to explain the gender discrepancy by citing previous work showing men tend to have worse ASR performance; however, 100% worse is much higher than peer studies (Sekkat et al., 2024; ElGhazaly et al., 2025; Attanasio et al., 2024).

3. Most first languages have statistically insignificant relative WER. This is due to those SG's not being represented by enough speakers in Fair-speech.

4. Native English speakers have negligibly higher worse-than-average WER, while several non-native speakers have statistically significantly

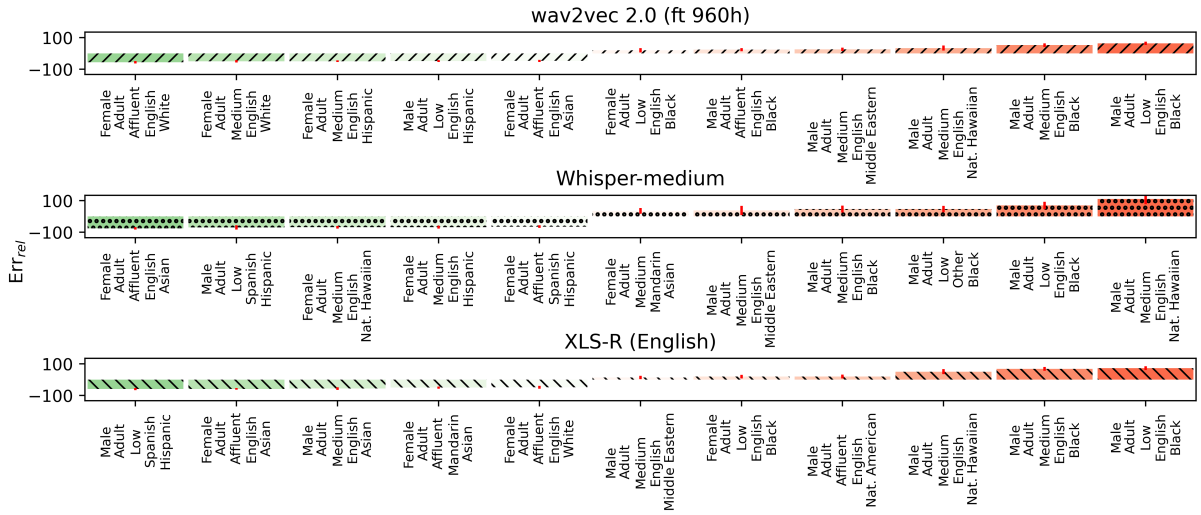

Fig. 8: Relative error of intersectional SG's in Fair-speech with least, greatest WER (conditional on sufficient speakers w.r.t. Eq. 1).

better-than-average WER. This stands in contrast to most peer studies (Feng et al., 2024; Fuckner et al., 2023; Sekkat et al., 2024; Ghorbani and Hansen, 2018). This is potentially an artifact of disregarding intersectionality of multivariate SG's.

5.4.1 Analysis of multivariate SG's can help Issue 1 is a warning that our "age" DV is probably improperly balanced w.r.t. other DV's. We reiterate that there is no good reason to analyze separate age groups of middle-aged adults; however, to investigate the hypothesis of imbalance in other DV's, we propose comparing only multivariate SG's where all DV's are the same except for age (see Section 5.4). Figure 6 shows that in doing so, the divergent performance w.r.t. age vanishes. That said, we note that the distribution for each class contains many outliers - this effect would likely diminish if we raised our threshold for number of speakers per multivariate SG. On the whole, this experiment supports our balancing hypothesis and supports the technique of multivariate comparison to avoid DV imbalance.

Issue 2 is similar to issue 1, and is similarly alleviated when considering only multivariate SG's differing only by gender (see Figure 6). As before, we note that the variance for both men and women is rather high - for some multivariate SG's, women have much higher WER than men, and vice versa. This further motivates in-depth analysis of individual multivariate SG's to uncover potential intersectional SG's with compounded ASR error.

We note that the ethnicity DV is the only one to deliver statistically significantly different results - for Black and White speakers. Based on this experiment, we conclude that being Black is associated with inferior ASR performance across every permutation of other DV's, while the opposite is true for being White. This is a stronger conclusion than what we were able to draw from Fig. 5.

5.4.2 Conditioning on all DV's is no silver bullet Conditioning on all DV's allowed us to draw strong conclusions about Black and White speakers; furthermore, we can reasonably rule out gender, age, or socio-economic background having broad impacts on ASR performance. However, we cannot draw statistically significant conclusions about many first languages (issue 3) even after aggregating uncommon languages into "other", due to persistent lack of representation across multivariate SG's. Thus, equipped with our conclusions about insignificance of gender, age, and socio-economic background, we propose removing conditions on those three DV's, considering only difference in first language and ethnicity. Furthermore, for first languages which we retained after aggregation, we note that apart from English, there is one primary ethnic group that covers nearly all speakers of that language (all native Mandarin speakers are of "Asian" ethnicity, almost all native Spanish speakers are "Hispanic", etc). Thus, we move to condition on ethnicity *only* for native English speakers.

Figure 7 allows us to get a better sense of bias against multivariate SG's defined by first language and ethnicity. We find once again that Black native English speakers are less well understood than every other class in every model. One added insight is that not all native English speakers are treated the same - White speakers are the best native English speakers understood by every model, though not statistically significantly so for all ethnicities. This provides a clearer picture than the previous experiment which showed that White speakers were statistically always better understood than the mean.

It also provides insight into issue 4, that non-native speakers are often better understood than native English speakers. When we condition on ethnicity, we find that it is only Black native English speakers that are worse understood, and that there is little statistical consensus regarding the relationship between being a native speaker and ASR error.

5.4.3 Intersectional multivariate SG's with compounded ASR error Finally, we can analyze the intersectional SG's which have the least and greatest WER, which we show in Figure 8. One surprising finding is that the worst performing group overall for `Whisper` is a Native Hawaiian group, which overall experienced much better treatment than Black speakers (Figure 7). One downside of this analysis, however, is that given the large difference in performance in extrema groups from the mean, fewer speakers are necessary to draw statistically significant conclusions based on SG (e.g. Eq. 1). Thus, such fine-grained analysis has heightened risk of drawing erroneous conclusions.

6. Discussion and outlook

The primary takeaway from this work is an exhortation to future studies on fairness in ASR to be as fastidious as possible designing their experiments. We underscore the importance of an intimate understanding of the datasets on which one is evaluating before designing experiments, tailoring experiments to that data, and being transparent about limitations that can be drawn therefrom.

We also encourage authors to clearly delineate the questions which they seek to answer. On the one hand, we can estimate how individual DV's affect performance of ASR systems, like in RQ 2 - this is primarily useful for understanding ASR systems from a computational level and could help steer future work towards disaffected SG's in particular. On the other hand, we can estimate which SG's defined by the intersection of multiple DV's are treated the absolute best and worst by ASR systems. This gives us greater sociological insight into the ramifications of unfair ASR - if a speaker belongs to such a class, they risk acute discrimination.

Furthermore, we encourage humility on behalf of researchers into ASR fairness in the face of statistical uncertainty. As we describe in this study, current fairness benchmarks suffer from lack of speaker diversity. Using overly broad SG's reduces the narrative power of their analysis, while using overly precise SG's silos speakers into groups without enough speakers, rendering conclusions statistically insignificant. With these two goalposts in mind, given the constraints of current ASR fairness benchmarks, **the conclusions we can draw from this type analysis remains limited**. Indeed reporting a result as statistically insignificant or sociologically broad isn't a problem (as long as this is duly noted); rather, it is a reflection of the reality of limitations of current corpora. This leads to the obvious recommendation to collect more data with high-quality annotations. However, this is both expensive and ethically fraught - Meyer et al. (2020) explicitly avoids children, while some countries forbid the analysis of ethnic minorities (fre, 1978).

It is important to note that the DV's included in the three benchmarks which we studied are themselves sources of potential bias. Had the benchmark designers decided to record different metadata, our results would reflect that, both in our ability to observe both the SG's which trigger unfairness, as well as the intersectional SG's which maximize unfairness. Future work in defining ASR fairness datasets should be in consultation with sociologists and phoneticians to determine which DV's, and SG's therein, are a) important in the larger context of social discrimination and b) likely to contribute to disparate ASR performance.

We encourage future study into linguistic or phonological mechanisms which are the actually underlying causal drivers of SG-level unfairness, beyond SG labels. Future work might eschew some of the pitfalls of relying on unbalanced and heterogeneous SG's advertised in this study by focusing on more rigorously defined measurements, such as dialect density measure (Koenecke et al., 2020) or speed of speech (Meng et al., 2022). Alternatively, unsupervised feature discovery has also been shown to uncover proxies for disadvantaged SG's in ASR without having to explicitly label them (Dheram et al., 2022; Alonzo-Canul et al., 2025). In other areas of ML, where fairness depends on sensitive attributes being evenly distributed in decision functions, there is work in automatically selecting attributes to include in fairness analysis (Pelegrina et al., 2023). This approach takes SG intersectionality into account by testing which combinations of attributes are related to an outcome variable.

One potential avenue for more precise fairness analysis is in conditional synthetic voice generation, such as (Sadok et al., 2025). This allows for two utterances from two different speakers to be *exactly* the same, apart from some parameters that are meant to mimic particularly DV's. For example, Masson and Carson-berndsen (2023) show that artificial generation simulates the patterns of non-native speakers well in ASR systems. This would address one major shortcoming of any current ASR fairness study, which is unable to truly isolate individual speaker characteristics while selecting from a small population of speakers.

7. Ethics statement

Collecting recordings of minority groups, particularly children, requires care to avoid revealing their identities. Fair-speech avoids children altogether. Furthermore, our work is meant to increase fairness in ASR; however, by focusing on a small number of datasets, we potentially overlook SG's which face ASR discrimination, thereby reinforcing it. In this case we didn't consider patients with adverse health conditions, for example, a group which has been studied extensively in ASR fairness research and no less deserving of attention than those highlighted in our study (Moro-Velázquez et al., 2019). By avoiding analyzing SG's for which benchmarks don't provide enough data, we are reinforcing the discrimination likely behind this unbalance in the first place.

8. Bibliography

1978. Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.

Mohammad Abushariah and Majdi Sawalha. 2013. The effects of speakers' gender, age, and region on overall performance of Arabic automatic speech recognition systems using the phonetically rich and balanced Modern Standard Arabic speech corpus. In *Proceedings of the 2nd Workshop of Arabic Corpus Linguistics*, Lancaster, UK.

Laura Alonzo-Canul, Benjamin Lecouteux, and François Portet. 2025. Vers l'apprentissage de modèles auto-supervisés de reconnaissance automatique de la parole plus équitables sans a priori démographique.

Frederic Aman, Michel Vacher, Solange Rossato, and Francois Portet. 2013. Speech recognition of aged voice in the AAL context: Detection of distress sentences. In *2013 7th Conference on Speech Technology and Human - Computer Dialogue (SpeD)*, pages 1–8, Cluj-Napoca, Romania. IEEE.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus.

Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 2024. Twists, Humps, and Pebbles: Multilingual Speech Recognition Models Exhibit Gender Performance Gaps.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias.

Lallouani Bouchakour and Mohamed Debyeche. 2022. Noise-robust speech recognition in mobile network based on convolution neural networks. *International Journal of Speech Technology*, 25(1):269–277.

Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist policies. *University of Chicago Legal Forum*, 1:139–167.

Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I.-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke. 2022. Toward Fairness in Speech Recognition: Discovery and mitigation of performance disparities. In *Interspeech 2022*, pages 1268–1272.

Alex DiChristofano, Henry Shuster, Shefali Chandra, and Neal Patwari. 2023. Global Performance Disparities Between English-Language Accents in Automatic Speech Recognition.

Michael Joseph Dino, Carla Leinbach, Gerald Dino, Ladda Thiamwong, Chloe Margalaux Villafuerte, Mona Shattell, Justin Pimentel, Maybelle Anne Zamora, Anbel Bautista, John Paul Vitug, Joyline Chepkorir, and Nerceilyn Marave. 2025. Smart Speakers for Health and Well-Being of Older Adults: A Mixed-Methods Review. *Healthcare*, 13(21):2772.

Hend ElGhazaly, Bahman Mirheidari, Nafise Sadat Moosavi, and Heidi Christensen. 2025. Exploring Gender Disparities in Automatic Speech Recognition Technology.

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84:101567.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying Bias in Automatic Speech Recognition.

James Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2019. An Intersectional Definition of Fairness.

Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2018. A comparative study of fairness-enhancing interventions in machine learning.

Marcio Fuckner, Sophie Horsman, Pascal Wiggers, and Iskaj Janssen. 2023. Uncovering Bias in ASR Systems: Evaluating Wav2vec2 and Whisper for Dutch speakers. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 146–151, Bucharest, Romania. IEEE.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2021. Investigating the Impact of Gender Representation in ASR Training Data: A Case Study on Librispeech. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 86–92, Online. Association for Computational Linguistics.

Shahram Ghorbani and John H. L. Hansen. 2018. Leveraging native language information for improved accented speech recognition. In *Interspeech 2018*, pages 2449–2453.

Katherine C. Hustad, Tristan J. Mahr, Phoebe Natzke, and Paul J. Rathouz. 2021. Speech Development Between 30 and 119 Months in Typical Children I: Intelligibility Growth Curves for Single-Word and Multiword Productions. *Journal of Speech, Language, and Hearing Research*, 64(10):3707–3719.

Wiebke Toussaint Hutiri and Aaron Ding. 2022. Bias in Automated Speaker Recognition. In *2022 ACM Conference on Fairness Accountability and Transparency*, pages 230–247.

Jongsuk Kim, Jaemyung Yu, Minchan Kwon, and Junmo Kim. 2025. FairASR: Fair Audio Contrastive Learning for Automatic Speech Recognition.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Ajinkya Kulkarni, Atharva Kulkarni, Miguel Couceiro, and Isabel Trancoso. 2024. Unveiling Biases while Embracing Sustainability: Assessing the Dual Challenges of Automatic Speech Recognition Systems. In *Interspeech 2024*, pages 4628–4632.

Linguistic Data Consortium. 2013. CABank English CallHome Corpus.

Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit.

Min Ma, Yuma Koizumi, Shigeki Karita, Heiga Zen, Jason Riesa, Haruko Ishikawa, and Michiel Bacchiani. 2024. FLEURS-R: A Restored Multilingual Speech Corpus for Generation Tasks.

Margot Masson and Julie Carson-berndsen. 2023. Investigating Phoneme Similarity with Artificially Accented Speech. In *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 49–57, Toronto, Canada. Association for Computational Linguistics.

Yen Meng, Yi-Hui Chou, Andy T. Liu, and Hung-yi Lee. 2022. Don't speak too fast: The impact of data bias on self-supervised speech models.

Josh Meyer, Lindy Rauchenstein, Joshua D. Eisenberg, and Nicholas Howell. 2020. Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6462–6468, Marseille, France. European Language Resources Association.

Laureano Moro-Velázquez, Jaejin Cho, Shinji Watanabe, Mark Hasegawa-Johnson, Odette Scharenborg, Heejin Kim, and Najim Dehak. 2019. Study of the Performance of Automatic Speech Recognition Systems in Speakers with Parkinson's Disease. pages 3875–3879.

Md Nayeem, Md Shamse Tabrej, Kabbojit Jit Deb, Shaonti Goswami, and Md Azizul Hakim. 2025. Automatic Speech Recognition in the Modern Era: Architectures, Training, and Evaluation.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 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.

Guilherme Dean Pelegrina, Miguel Couceiro, and Leonardo Tomazeli Duarte. 2023. A statistical approach to detect sensitive features in a group fairness setting.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Ana Rodrigues, Rita Santos, Jorge Abreu, Pedro Beça, Pedro Almeida, and Sílvia Fernandes. 2019. Analyzing the performance of ASR systems: The effects of noise, distance to the device, age and gender. In *Proceedings of the XX International Conference on Human Computer Interaction*, Interacción '19, pages 1–8, New York, NY, USA. Association for Computing Machinery.

Sandra Rojas, Elaina Kefalianos, and Adam Vogel. 2020. How Does Our Voice Change as We Age? A Systematic Review and Meta-Analysis of Acoustic and Perceptual Voice Data From Healthy Adults Over 50 Years of Age. *Journal of Speech, Language, and Hearing Research*, 63(2):533–551.

Samir Sadok, Simon Leglaive, Laurent Girin, Gaël Richard, and Xavier Alameda-Pineda. 2025. AnCoGen: Analysis, Control and Generation of Speech with a Masked Autoencoder.

Chloe Sekkat, Fanny Leroy, Salima Mdhaffar, Blake Perry Smith, Yannick Estève, Joseph Dureau, and Alice Coucke. 2024. Sonos Voice Control Bias Assessment Dataset: A Methodology for Demographic Bias Assessment in Voice Assistants. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15056–15075, Torino, Italia. ELRA and ICCL.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 59–68, New York, NY, USA. Association for Computing Machinery.

Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Rachael Tatman and Conner Kasten. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *Interspeech 2017*, pages 934–938. ISCA.

Irina-Elena Veliche, Zhuangqun Huang, Vineeth Ayyat Kochaniyan, Fuchun Peng, Ozlem Kalinli, and Michael L. Seltzer. 2024. Towards measuring fairness in speech recognition: Fair-Speech dataset.

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, pages 1–7, New York, NY, USA. Association for Computing Machinery.

Rebecca Wald, Jessica Taylor Piotrowski, Johanna M.F. Van Oosten, and Theo Araujo. 2024. Who are the (Non-)Adopters of Smart Speakers? A Cross-Sectional Survey Study of Dutch Families. *Tijdschrift voor Communicatiewetenschap*, 52(1):4–28.

Angelina Wang, Vikram V. Ramaswamy, and Olga Russakovsky. 2022. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. In *2022 ACM Conference on Fairness Accountability and Transparency*, pages 336–349.

Wenbin Wang, Yang Song, and Sanjay Jha. 2024. GLOBE: A High-quality English Corpus with Global Accents for Zero-shot Speaker Adaptive Text-to-Speech. In *Interspeech 2024*, pages 1365–1369. ISCA.

Rumei Yang, Shiying Gao, and Yun Jiang. 2024. Digital divide as a determinant of health in the U.S. older adults: Prevalence, trends, and risk factors. *BMC Geriatrics*, 24(1):1027.

Yuanyuan Zhang, Yixuan Zhang, Bence Halpern, Tanvina Patel, and Odette Scharenborg. 2022. Mitigating bias against non-native accents. In *Proc. Interspeech 2022*, pages 3168–3172.

9. Language Resource References

Veliche, Irina-Elena and Huang, Zhuangqun and Kochaniyan, Vineeth Ayyat and Peng, Fuchun and Kalinli, Ozlem and Seltzer, Michael L. 2024. *Towards Measuring Fairness in Speech Recognition: Fair-Speech Dataset*. arXiv, arXiv:2408.12734. PID https://ai.meta.com/datasets/speech-fairness-dataset/.

A. Regression with constituent DV's

Another technique which has been used in the literature to estimate the effect of individual DV's on ASR performance is fitting a regression to predict ASR system's error on each speaker based on their constituent DV's (Sekkat et al., 2024; Tatman, 2017; DiChristofano et al., 2023). As in the previous case, it is imperative to fit this regression based on mean speaker performance rather than overall utterances in order to avoid biasing it towards individual speakers. The simplest form, regarding only univariate models and assuming categorical SG's, takes the following form:

$$\text{WER avg.}(D) = \sum_{DV^i} \sum_{SG \in DV^i} \alpha_{SG} \cdot \mathbb{1}_{(S=SG)} \quad (7)$$

where we expand DV^i to include the everything-SG (to simulate a bias term α_0). Sekkat et al. (2024) then goes on to define a multivariate model, which takes the intersection of SG's into account:

$$\text{WER avg.}(D) = \left(\sum_{DV^j} ... \sum_{DV^k} \sum_{SG_j \in DV^j} ... \sum_{SG_k \in DV^k} \right) \alpha_{SG_j,...,SG_k} \cdot \mathbb{1}_{(S=SG_j,...,SG_k)} \quad (8)$$

which is the semantic equivalent of Section 5.4. However, we consider relative SG-level WER to be a more intuitive measure; therefore, we focus on this for the remainder of the study.