MODEL-BASED APPROACH FOR MEASURING THE FAIRNESS IN ASR

Zhe Liu, Irina-Elena Veliche, Fuchun Peng

Meta AI, Menlo Park, CA, USA

INTRODUCTION

Automatic speech recognition (ASR) systems are getting better and better with the advent of new technologies. However, the issue of *fairness* arises when these tools do not perform equally well for all subgroups of the population.

In most of previous research studies on measuring fairness in ASR, WER was computed per each subgroup (e.g. black speakers versus white speakers) and from the comparison of these WER numbers, conclusions were drawn on whether significant disparities exist among these subgroups of interest. Although this is a very simple way for fairness measurement, there are several open questions that have not yet been properly addressed in these analyses.

- First, how to effectively control the *confounding* factors which may affect the measured results but are not of primary interest? For example, we need to deal with any unbalanced gender or age distribution of speakers in a racial disparity study; otherwise, it would be difficult to tell whether any WER gap among different racial groups is due to the race factor or any confounding factor of gender or age.
- Secondly, how to appropriately take into account speaker-level effect on measured WERs and handle unobserved *heterogeneity* across different speakers?
- Third, how to efficiently trace the source of any WER gap among different subgroups, that is, does such disparity mainly come from phonetic, phonological, prosodic characteristics, or grammatical, lexical, semantic characteristics, or both?

In this paper, we present a model-based approach to better measure the fairness issue in ASR and study any performance disparities across different subgroups of our interest. In particular, we introduce *mixed-effects Poisson regression*, treating utterance-level word errors as the regression response, logarithm number of words in the reference text as an offset, speaker identification as a random effect, subgroup label of interest and any other explanatory or confounding variables as fixed effects. In particular, our proposed method prevents underestimating the standard errors and avoids drawing false positive conclusions on non-fairness.

Simulation on confounding factor experiment

Confounding Rate		**Baseline**		**Model-Based**	
within Case	*within Control*	*Mean Ratio*	*% False Positive*	*Mean Ratio*	*% False Positive*
50%	50%	1.000	4.9%	1.000	4.7%
60%	40%	1.021	12.1%	1.001	5.8%
70%	30%	1.041	29.8%	1.000	5.4%
90%	10%	1.084	83.3%	1.001	5.1%

Simulation on speaker effect experiment

Speaker Effect		**Baseline**		**Model-Based**	
Num of Speakers	*Standard Deviation*	*Mean Ratio*	*% False Positive*	*Mean Ratio*	*% False Positive*
500	0.2	1.000	8.0%	1.000	4.8%
500	0.4	1.001	14.9%	1.001	4.5%
100	0.2	1.000	16.6%	1.000	5.0%
100	0.4	0.999	42.6%	0.999	5.2%

METHODS

Poisson regression serves as an appropriate approach to model rate data, where the rate is a count of events (e.g. word errors in our use case) divided by some measure of that unit's exposure (e.g. number of words in the reference). An *offset* variable is needed to scale the modeling of the mean parameter in Poisson regression with a log link.

More specifically, to measure the effect of factor $f(\cdot)$ on WER results across l different subgroups, the vanilla Poisson regression model is described as follows:

$$C_s \overset{\text{i.i.d.}}{\sim} Poisson(\lambda_s) \tag{1}$$

$$\log(\lambda_s) = \log(N_s) + \mu_{f(s)} \tag{2}$$

where C_s is the count of word errors (sum of insertion, deletion, and substitution errors), λ_s is the Poisson (mean) parameter, N_s is the number of words in the reference text for the sth utterance in the evaluation dataset, and $\mu_{f(s)}$ refers to the factor effect corresponding to the subgroup of $f(s)$.

It is natural and flexible to extend the vanilla Poisson regression model (1) (2) to include additional explanatory or confounding covariates, which can be utilized to capture effects of confounding variables on WERs among different subgroups:

$$\log(\lambda_s) = \log(N_s) + \mu_{f(s)} + \theta^T x_s \tag{3}$$

Here, x_s represents the vector of any explanatory variables in the regression model and θ refers to the coefficient parameter vector that shall be learned. For example, in a racial disparity analysis, we would want to add the gender or age information of speakers to the regression model in order to control any confounding effects.

Block-structured evaluation data arises naturally in any real-world speech recognition applications. In particular, utterances from the same speaker could share common correlated features (e.g. accent of speaker), and thus analyses that assume independence of these observations will be inappropriate. The use of random effect is one usual and convenient way to model such structure.

A mixed-effects Poisson regression is a model containing both fixed effect and random effect. Regarding the fairness measurement of speech recognition accuracy among different subgroups of the factor $f(\cdot)$, we describe the model in detail as follows:

$$r_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{4}$$

$$C_{ij} \,|\, \lambda_{ij} \overset{\text{i.i.d.}}{\sim} Poisson(\lambda_{ij}) \tag{5}$$

$$\log(\lambda_{ij}) = \log(N_{ij}) + \mu_{f(i)} + r_i + \theta^T x_{ij} \tag{6}$$

where the utterance-level index of subscription notation ij represents the jth utterance from the ith speaker, r_i denotes the speaker-level random effect that is independently sampled from a Gaussian distribution with mean 0 and variance σ^2 which is learnable. Note that any C_{ij} and $C_{ij'}$ are no longer independent for $j \neq j'$ since they are observed from the same speaker i, while any $C_{i\cdot}$ and $C_{i'\cdot}$ are still independent for $i \neq i'$ since they are observed from different speakers. Also, we use $\mu_{f(i)}$ to denote the fixed effect for the factor $f(\cdot)$ of primary interest, since typically it is at speaker level.

SIMULATION EXPERIMENTS

We conduct simulation experiments to show that the proposed mixed-effects Poisson regression could properly address the problems of confounding factor and speaker effect in ASR fairness measurements.

From the tables on the left hand side, we observe high false positive rates for the baseline method. Instead, the model-based approach always results in approximate 5% false positive rate. This demonstrates that it can successfully deal with confounding factor and speaker effect and is superior than the traditional baseline method.

REAL DATA EXPERIMENTS

Apply the proposed mixed-effects Poisson regression on real-world speech datasets for fairness investigation.

- *LibriSpeech*. A widely used voice dataset which consists of 960 hours transcribed training utterances. The evaluation dataset has the splits of *Test-Clean* from 40 speakers and *Test-Other* from 33 speakers.
- *Voice Command*. De-identified dataset collected using mobile devices through crowd-sourcing from a data supplier for ASR. The participants are instructed to say voice commands on calling, playing music, etc. The evaluation set contains around 18K utterances from 95 speakers.

For the *LibriSpeech* data, we study the ASR fairness on *gender*, that is, we would like to test whether there exists statistical significance on the WER ratio between male speakers and female speakers.

The **baseline** approach, which is widely used in practice, computes the ratio of empirical WER from male speakers group over the empirical WER from female speakers group. The bootstrap method is applied to compute the 95% CI of the ratio. For **model-based** approach, we fit a mixed-effects Poisson regression based on (4) (5) (6) with gender as the fixed effect and speaker label as a random effect.

The baseline method leads to statistically significance claims on both *Test-Clean* and *Test-Other* sets, and interestingly, their conclusions are actually opposite. Specifically, on *Test-Clean* split the baseline method shows that male speakers group has significant lower WER compared to female speakers group, while on *Test-Other* split, male speakers group has significant higher WER compared to the group of female speakers. On the other hand, the model-based approach does not claim any significant results on both splits. This makes sense since numbers of speakers in both splits are quite small, which results in high variance estimation that does not lead to statistically significant results. Thus utterances from more speakers are needed to reduce the standard errors and draw a more sound conclusion.

	Baseline		**Model-Based**	
LibriSpeech Dataset	*WER Ratio*	*Confidence Interval*	*WER Ratio*	*Confidence Interval*
Test-Clean	0.86	(0.76, 0.97)	0.88	(0.67, 1.14)
Test-Other	1.34	(1.23, 1.46)	1.28	(0.93, 1.76)

We also investigate ASR fairness on gender for *Voice Command* dataset. The baseline method does not claim that the WER on male speakers group is statistically significantly higher than the WER of female speakers group, but it is very close. The model-based method clearly does not lead to significant result, due to the relatively small number of speakers in each group.

	Baseline		**Model-Based**	
Voice Command Dataset	*WER Ratio*	*Confidence Interval*	*WER Ratio*	*Confidence Interval*
Test	1.08	(0.99, 1.20)	1.15	(0.78, 1.72)

CONCLUSION

We introduce mixed-effects Poisson regression to better measure and interpret any WER difference among subgroups of interest. The presented method is very flexible and can effectively address the open problems of how to control the confounding factors, how to handle unobserved heterogeneity across speakers, and how to trace the source of any WER gap among different subgroups.