Investigating the Impact of Gender Representation in ASR Training Data: a Case Study on Librispeech

Mahault Garnerin LIDILEM & LIG
Solange Rossato LIG
Laurent Besacier LIG
Univ. Grenoble Alpes, CNRS, Grenoble INP
FR-38000 Grenoble, France
firstname.lastname@univ-grenoble-alpes.fr

Abstract

In this paper we question the impact of gender representation in training data on the performance of an end-to-end ASR system. We create an experiment based on the Librispeech corpus and build 3 different training corpora varying only the proportion of data produced by each gender category. We observe that if our system is overall robust to the gender balance or imbalance in training data, it is nonetheless dependant of the adequacy between the individuals present in the training and testing sets.

1 Introduction

As pointed out by Hovy and Spruit (2016) in their positional paper on the social impact of NLP, discriminatory performance could be the result of several types of biases. The roots of socio-technical biases in new technology could be situated in its very design, the selection of the data used for training (Garg et al., 2018; Kutuzov et al., 2018), the annotation process (Sap et al., 2019), the intermediary representations such as word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017) or in the model itself.

Gender bias in ASR systems, defined as a systematically and statistically worse recognition for a gender category is still a working topic (Feng et al., 2021). Pioneer work from (Adda-Decker and Lamel, 2005) found better performance on women's voices, while a preliminary research on YouTube automatic caption system found better recognition rate of male speech (Tatman, 2017) but no gender-difference in a follow-up study (Tatman and Kasten, 2017). Recent work on hybrid ASR systems observed that gender imbalance in data could lead to decreased ASR performance on the gender category least represented (Garnerin et al., 2019). This last study was conducted on French broadcast data in which women account for only 35% of the speakers. If systematic, this performance difference could lead to less indexing of media resources featuring female speech and contribute to the invisibilisation of women and women speech in public debate[1] and history (Adamek and Gann, 2018). Such results would also fall into the category of allocational harms, following the typology proposed by Barocas et al. (2017) and Crawford (2017), because women are more likely to be less represented in corpora (Garnerin et al., 2020), making all technologies relying on speech recognition less accessible for them. It could also result in representational harm such as the maintenance of the stereotype of inadequacy between women and technology.[2] But as other characteristics such as the speaker role, (i.e. his or her ability to produce professional speech) could explain some performance variations, we propose in this paper to address the question of ASR systems' robustness to gender imbalance in training data. As data is now the starting point of every system, we know that the quality of a system depends on the quality of its data (Vucetic and Obradovic, 2001; He and Garcia, 2009). To tackle this question, we work with the Librispeech corpus, widely used in the community and based on audio books recordings. To evaluate the impact of gender imbalance in training data on our system performance, we proceed as follows: we first test the robustness of our model against the randomness introduced at training by the weight initialization stage. We then evaluate

[1] https://www.newyorker.com/culture/cultural-comment/a-century-of-shrill-how-bias-in-technology-has-hurt-womens-voices

[2] see for example, this news report on decreased performance for female speaker in built-in GPS, in which the VP of voice technology stated "many issues with women's voices could be fixed if female drivers were willing to sit through lengthy training... Women could be taught to speak louder, and direct their voices towards the microphone" https://techland.time.com/2011/06/01/its-not-you-its-it-voice-recognition-doesnt-recognize-women/

Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing, pages 86–92

the impact of speakers selection in training data on model performance. We compare the obtained results to the impact observed when changing overall gender representation in training data. Finally we observe the behavior of our model when trained on mono-gender corpora.

We validate our model robustness against the impact of model seed and observe that overall system is quite robust to gender balance variation. We note that the random factor introduced in the selection process of speakers for the training set seems to have a statistically significant impact on performance. We argue that our model, whereas robust to gender representation variability, is strongly dependent on the individuals present in the training set, which questions the pertinence of gender as a category of analysis in ASR and advocate for a return to a more incorporated conception of language.

2 End-to-end model of Automatic Speech Recognition

For the last decade, the traditional ASR models, based on HMM-GMMs have been coexisting with hybrid models (HMM-DNNs) (Mohamed et al., 2012; Dahl et al., 2012) and for the latest couples of years with end-to-end systems. The former acoustic, pronunciation and language models, made explicit by different modules in the final system, are now collapsed into one big architecture mapping directly the audio signal to its transcription. Since speaker adaptation has been integrated into the entire training process of end-to-end models, we are expecting the gender imbalance within training data to be extrapolated by this kind of systems, resulting in gender-differentiated performance.

2.1 Original data set

We used the Librispeech data set (Panayotov et al., 2015) to perform our experiments. The original training data set contains a total of 5466 books read by 2338 US English speakers. 2671 books are read by female speakers and 2795 by male speakers. As we decide to use the gender terminology over the sex one, we acknowledge that staying within these binary categories is reductive. However, as there is no mention of non-binary speakers in our data sets and believing that the audit of discriminatory performance on non-binary people calls for a thought-through methodology, we stayed within the binary matrix. We are nonetheless aware of the limitations that comes with this choice.

Data set	F	M	Total
train original	2671	2795	5466
wper30	1145	2671	3816
wper50	1908	1908	3816
wper70	2671	1145	3816
test-clean	49	38	87

Table 1: Composition of the different training and evaluation data sets. Numbers reported are numbers of books read by men and women.

The Librispeech corpus comes with two testing sets : test-clean and test-other. The test-clean contains 87 books read by 40 speakers. 49 books are read by women and 38 by men. The test-other set contains 90 books read by 33 speakers, in which 44 books are read by women and 46 by men. The test-clean includes speakers which obtained the best WER according to the results of the WSJ model's transcripts and the speakers left were put in the test-other data set. In this work, analyses are conducted on the test-clean set.

We decide to work at the book granularity. Meaning each point of measure is the WER obtained on a particular book. There is no speaker overlap between train and test sets. For the sake of readability, when we report WER results for male and female speakers, we actually refer to WER results obtained for books read by male or female speakers.

2.2 Controlled data sets

Librispeech being gender balanced by design, we recreated 3 training data sets in which 30%, 50% or 70% of the books were read by women, in order to observe the impact of gender balance on performance. We called the resulting training sets: wper30, wper50 and wper70. To assure comparability, the overall number of books (N=3816) is the same for each training set. The common part between each data set is maximised : the 30% of books read by women in wper30 are also present in the wper50 and wper70 data sets. The same applies to books read by men. We then trained a system with each one of them.

2.3 Model

We trained our systems with the ESPnet toolkit (Watanabe et al., 2018) and used a state of the art model based on an already existing recipe: our model is an attentional encoder-decoder model, with a 5-layer VGG-BLSTM encoder and a 2-layer

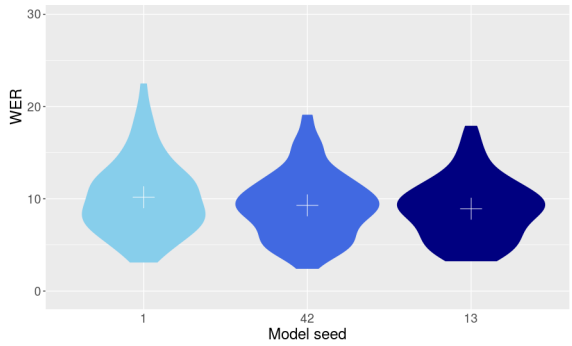

Figure 1: WER distributions on test-clean testing set by model seeds. White crosses represent the mean value of each distribution and color represent each model. Training done on wper50 partition.

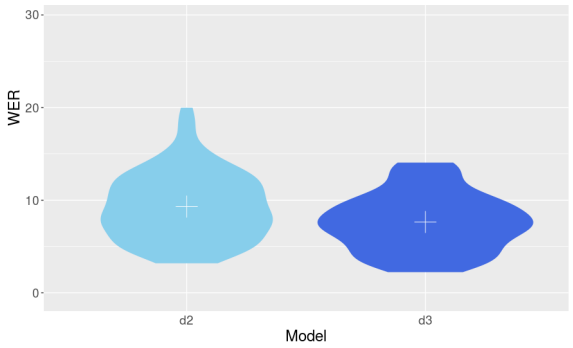

Figure 2: WER distributions on test-clean testing set by data seeds. Both systems have a 50-50% men/women training set. White crosses represent the mean value of each distribution and color represent each model. Training done on wper50 partition.

decoder. The encoder and decoder layers have 1024 hidden units and the output vocabulary is of 5,000 subwords generated through byte pair encoding. We used the PyTorch backend for ASR training and decoding was performed using both an RNN LM trained on the Librispeech text corpus and the joint decoding combining attention-based and CTC scores of the ASR model (CTC weight=0.5, LM weight=0.7).

With this configuration we obtained (with a model learnt on the full train set) a mean WER of 4.2% on the test-clean data set and a mean WER of 14.3% on the test-other set. Reported results on the ESPnet repository were of 4.0% on test-clean and 12.7% on test-other with a similar configuration.

2.4 Statistical testing of WER results

To assess the existence of a statistically significant impact on performance of the different conditions tested, we chose non-parametrical tests, considering our WER distributions do not follow a normal distribution. We used the Wilcoxon Rank Sum test (also known as Mann-Whitney test) and its generalisation to more than 2 samples, the Kruskall-Wallis test (Wilcoxon et al., 1963). Both tests estimate the probability of the WER distributions to be samples of the same population. We set our confidence level to 99% ($\alpha = 0.01$).

3 Impact of the model seed

Our hypothesis that systems might be impacted by a gender imbalance in training data is based on the fact that systems are deeply dependent on the data they are trained on (Vucetic and Obradovic, 2001; He and Garcia, 2009). In order to control that the behaviors we observe are only due to the data variation, we conduct a first experiment where we test the robustness of our model to the seed variability at training. To do so, we train three models with the wper50 (gender-balanced) training set, changing only the model seed. Obtained WER distributions are represented in Figure 1. When performing the Kruskall-Wallis test, no statistical significant difference is observed between the 3 distributions (p-value = 0.17). The same observation is made when comparing the models two by two. We conclude that our model is robust to the randomness introduced at the initialisation stage.

4 Impact of the training data (data seed)

We believe that gender is an attribute of the speaker and that speaker's gender variability goes beyond gender statistics. Following Judith Butler's theory on the performativity of gender (Butler, 1988, 2011), we assume that gender is not expressed in the same way amongst speakers. The intrinsic variability of gender indexing (Ochs, 1992) leads us to consider that two people sharing the same gender "label" will not be interchangeable in a data set.

In order to test this hypothesis, we created two other training sets with a 50-50 men/women balance but with a different random seed for the shuffle and selection process for these training corpora. We refer to this random element as the "data seed". We call the two models d2 and d3 (data seeds values were chosen arbitrarily). We obtained the distributions presented in Figure 2. Wilcoxon rank sum test is statistically significant between the two distributions (W=4771.5; p-value=0.003). Model d2 obtains a mean WER of 9.31% and model d3 a

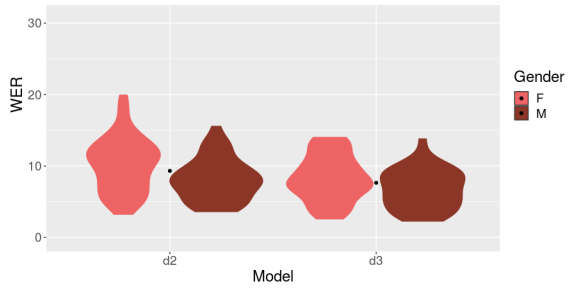

Figure 3: WER distributions on test-clean testing set by data seeds. Black dots represent the mean value of each distribution regardless of gender categories. Training done on wper50 partition.

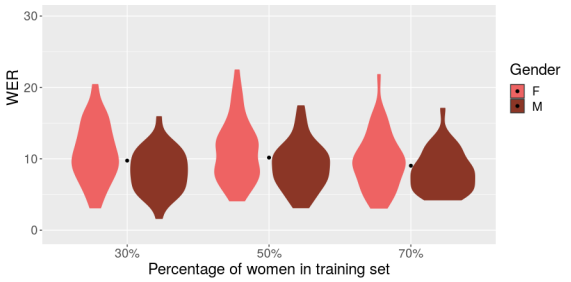

Figure 4: WER distributions on test-clean testing set by gender for the 3 models trained with 30%, 50% or 70% of books read by women in the training set. Black dots represent the mean value of each distribution regardless of gender categories.

mean WER of 7.65%. We argue that the data exhaustiveness is thus a strong factor of performance variation.

When looking at the performance obtained by gender categories (see Figure 3), we also observe distinct behaviors between the two models. WER for books read by male (8.14%) and female (10.2%) speakers are statistically different in our model d2 (W=1240; p-value=0.008). This effect is not found in model d3 (W=1173; p-value=0.038), although a difference of almost 2 points is also observable between the mean WER for men (6.80%) and women (8.31%). There is trend to obtain slightly better WER results for male speakers, with an average difference of 1.7 percentage point. The performance difference between the gender categories is thus of the same order as the difference between our models d2 and d3.

5 Gender balance and performance

In this experiment we try to evaluate the impact of gender representation on the performance. To do so, we trained 3 ASR systems, with our 3 different training sets presented in Section 2.2. Results are reported in Table 2. Overall WERs are of 9.7% respectively 10,2% and 9.0% for our 3 conditions (training set with 30% of books read by women, respectively 50% and 70%). We note a decrease in WER performance for wper50 that could be explained by a different speakers selection for training, as we observed in the previous Section 4. However, no statistical difference is observed between these 3 conditions (p-value = 0.14).

A quick look at our WER distributions by gender category shows that the performance obtained for women are generally worse than the one obtained for men (see Figure 4). This difference is statistically significant (p-value = 0.003) when our training set contains only 30% of books read by women and p-value increases until it exceeds our alpha risk (p-value = 0.04 for wper50 and p-value = 0.10 for wper70). We can argue that an under-representation of a gender category leads to a higher error rate, but the same trend is not observable for male speakers. Surprisingly, when training set contains 70% of books read by women, there is no significant difference between WER obtained for male and female speakers. Even if it is not statistically significant, the trend observed in Section 4 holds because with 70% of female speech in training data, we still observe better WER results for men.

6 What about mono-gender models?

Our overall system performance seems to be robust to the variation in gender representation in training data. In wper30 model we observe a statistically significant gender difference in WER. The better WER results for male speakers are expected as they are more represented in training data. This is not the case for wper70 model, where we expected better results for women. Therefore we trained male-only and female-only models to analyse extreme behaviors. We maximized the size of our training set, reaching a book count in these mono-gender systems of 2671. Hence, it is worth noting that the size of training data in these systems is smaller than the size of training data for our wper models.

Overall WER for the male-only model is of 12.3% and 11.7% for the female-only one without any statistically significant difference. In the male-only model, WER distributions are statistically different by gender category (p-value $< 10^{-6}$), with an average WER of 9.11% on books read by men and of and of 14.7% on books read by

Model	Gender	test-clean
wper30	F	10.9%
	M	8.3%
	all	**9.7%**
wper50	F	11.0%
	M	9.1%
	all	**10.2%**
wper70	F	9.6%
	M	8.3%
	all	**9.0%**

Table 2: Mean WER by gender obtained on the Librispeech test-clean data set for the 3 models trained with 30%, 50% or 70% of books read by women in the training set.

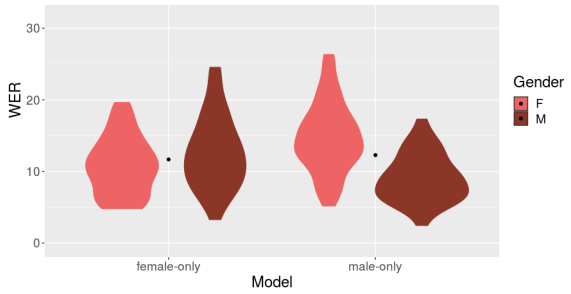

Figure 5: WER distributions on test-clean testing set by gender for our two mono-gender models. Black dots represent the mean value of each distribution regardless of gender categories.

women. But this is not the case for the female-only model (p-value = 0.114 ; WER(F) = 10.9% and WER(M)=12.7%. At last, when we are in a mono-gender configuration with only women-read books at training, we reverse the trend of better WER results for male speakers, but without reaching statistical significance. It seems that an over-representation of women is better suited to the task in our experimental settings.

7 Discussion & Conclusion

It is a common-sense claim to state that all gender categories need to be represented in a training corpus of an ASR system in order to be able to transcribe speech regardless of the user's gender. We expected to find that the performance obtained on each gender category was dependent of their representation in training data. However, if we select individuals while maintaining a balanced gender distribution (see Section 4), we obtain a significant difference in performance of around 1.7 percentage point. It is possible that these differences in performance, between systems and between genders, will not be found for other test corpora, because more than the selection of individuals present at training, it is the "proximity" between voices in the training and test sets that may explain these observed differences. When varying the percentage of female-read books in training sets, we find that the global performance keeps the same range of accuracy, without any statistical significance. As individuals also change when varying the proportion of men and women in training data, we expected our WER distribution to vary accordingly. However, it is worth noting that our three data sets always include the 30% of women of the wper30 set and the 30% of men of the wper70 set. Surprisingly, for the three systems studied, both the gendered proportion and the individuals change without inducing significant differences in overall performance.

We observe that the expected impact is not in terms of overall performance but in terms of performance by gender. There is a higher error rate for female speakers when the system is mostly trained on male speakers but the significance of this difference between men and women decreases slowly as we raise the quantity of female-read books in the training set. However, we do not observe the inverse trend: only with the mono-gender system trained with women voices only, do we achieve better WER results for women than for men, even if this trend is not significant. All in all, we cannot conclude that the gender distributions in the training data have a strong influence on the WER results. While it appears that men's voices are generally better recognised, it seems that increasing the proportion of women's voices in the training corpus helps to reduce gender-differentiated performance, while ensuring the same level of overall performance.

From this study performed on Librispeech, it appears that i) the selection of individuals in the training corpus, ii) the gender distribution with extreme variations and iii) the train/test corpus match have a significant impact on system performance. In this very controlled context of speech production, the gender variation seems to be negligible compared to the individual variation. We believe gender demographics are not enough to ensure the same level of performance on both gender groups. According to our results, it seems that an over-representation of female voices improves recognition of women voices without decreasing overall

performance. Further research is needed to disentangle the effects of gender representation in voice and data and the performance of ASR systems. If considering the gender balance in training data is a starting point for fairer systems, trying to quantify the intra-variability of our training sets to estimate a measure of adequacy with our test data appears as a strong lead for future work. We plan on working on acoustic measures such as fundamental frequency and speech rate to assess something that could be named “voice variability cover" and try to finally get out of the binary sex-matrix.

References

Anna Adamek and Emily Gann. 2018. Whose artifacts? whose stories? public history and representation of women at the canada science and technology museum. *Historia Crítica*, 68:47–66.

Martine Adda-Decker and Lori Lamel. 2005. Do speech recognizers prefer female speakers? In *Proceedings of the 9th European Conference on Speech Communication and Technology*, INTERSPEECH 2005, pages 2205–2208, Lisbon, Portugal. ISCA.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *Proceedings of SIGCIS Conference*, Philadelphia, PA, USA.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS 2016, pages 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Judith Butler. 1988. Performative acts and gender constitution: An essay in phenomenology and feminist theory. *Theatre journal*, 40(4):519–531.

Judith Butler. 2011. *Bodies that matter: On the discursive limits of sex*. Routledge, London.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Kate Crawford. 2017. The trouble with bias. Keynote at the 31st Annual Conference on Neural Information Processing Systems, NIPS 2017, Long Beach, CA, USA.

George E. Dahl, Dong Yu, Li Deng, and Alex Acero. 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. (Submitted to INTERSPEECH 2021).

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 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*, AI4TV ’19, pages 3–9, Nice, France. ACM.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2020. Gender representation in open source speech resources. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6599–6605, Marseille, France. European Language Resources Association.

Haibo He and Edwardo. A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Abdel Rahman Mohamed, George. E. Dahl, and Geoffrey E. Hinton. 2012. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):14–22.

Elinor Ochs. 1992. Indexing gender. In Alessandro Duranti and Charles Goodwin, editors, *Rethinking Context: Language as an interactive phenomenon*, pages 335—350. Cambridge University Press.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, South Brisbane, QLD, Australia. IEEE.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

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 *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, INTERSPEECH 2017, pages 934–938, Stockholm, Sweden. ISCA.

Slobodan Vucetic and Zoran Obradovic. 2001. Classification on data with biased class distribution. In *Machine Learning: ECML 2001*, pages 527–538. Springer Berlin Heidelberg.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-end speech processing toolkit. In *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, INTERSPEECH 2018, pages 2207–2211, Hyderabad, India. ISCA.

Frank Wilcoxon, SK Katti, and Roberta A Wilcox. 1963. *Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test*. American Cyanamid Company, Pearl River, NY, USA.