AI Safety

September 2024

Analysis of Progress in Speech Recognition Models

Miguel A. Peñaloza

penaloza@cicese.edu.mx

CARRERAS CON
IMPACTO

INTRODUCTION

This project initially aimed to estimate the progress of speech recognition models by means of scaling laws (Hendricks, 2024). Through the variables of **FLOPS (Floating Point Operations), number of model parameters, size of the training sample in hours, architecture of the neural networks and WER metrics (Word Error Rate).**

INTRODUCTION

The FLOPS (number of floating point operations) were estimated using the methodology number two reported by Sevilla et al. (2022):

(training time) X (# de cores) X (# peak FLOPS) X (utilization rate).

INTRODUCTION

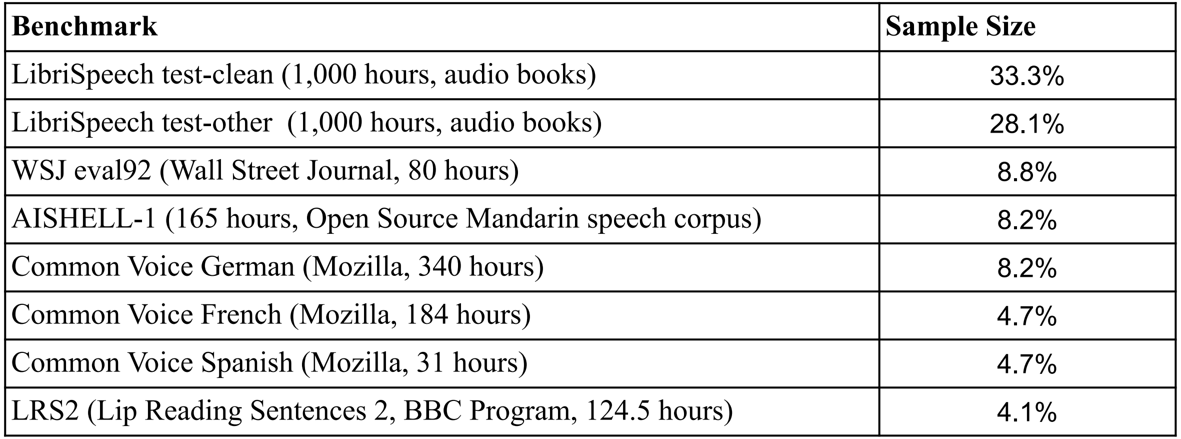
WER (Word Error Rate) metric the **number of errors** is calculated as the sum of **substitutions (S), insertions (I) and deletions (D) divided by the total number of words (N) and multiplied by 100:**

$$WER = \frac{I+D+S}{N} \times 100$$

Ec 1. WER (Word Error Rate) The lower the WER metric, the better the performance of the model since the error rate is lower for more details we suggest consulting NithyaKalyani & Jothilakshmi, (2019).

METHODS

Step 1: Compilation and construction of a research **dataset** from the Browse State of The Art repository in the area of **Speech recognition.** Obtaining a **sample size of 171 speech recognition models.**



Benchmark	Sample Size
LibriSpeech test-clean (1,000 hours, audio books)	33.3%
LibriSpeech test-other (1,000 hours, audio books)	28.1%
WSJ eval92 (Wall Street Journal, 80 hours)	8.8%
AISHELL-1 (165 hours, Open Source Mandarin speech corpus)	8.2%
Common Voice German (Mozilla, 340 hours)	8.2%
Common Voice French (Mozilla, 184 hours)	4.7%
Common Voice Spanish (Mozilla, 31 hours)	4.7%
LRS2 (Lip Reading Sentences 2, BBC Program, 124.5 hours)	4.1%

METHODS

Due to drawbacks in the construction of the dataset, for example that most of the researches consulted **do not report the computation used (FLOPS), nor the parameters to estimate them,** it was decided to continue the analysis **using only the WER metric and architecture of the neural networks.**

RESULTS

- Common Voice French (WER 10.5%) y Common Voice Spanish (WER 7.5%).
- **LibriSpeech Test Clean (WER 3%) y Wall Street Journal (WSJ 92, WER 3.3%).**

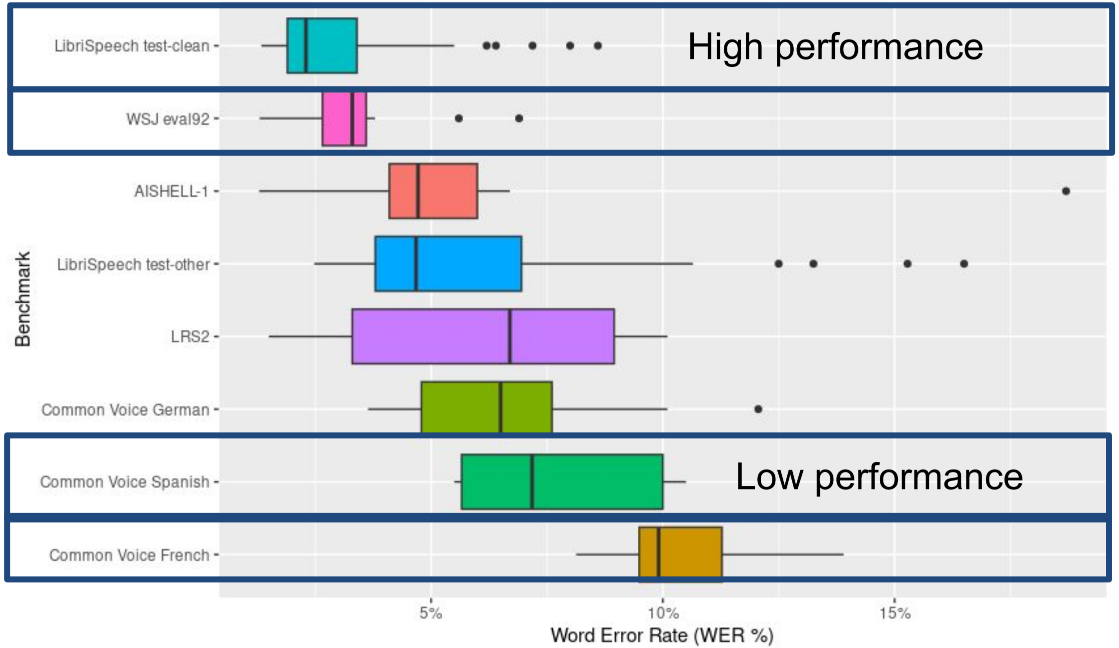


Figure 3. Distribution of Word Error Rate (WER%) in different speech recognition benchmarks.

RESULTS

- Convolutional Neural Network (WER 6%) and (Hidden Markov Model, WER 5.6%).

- **Transformer (WER 3.17%) y Conformer (WER 2.67%).**

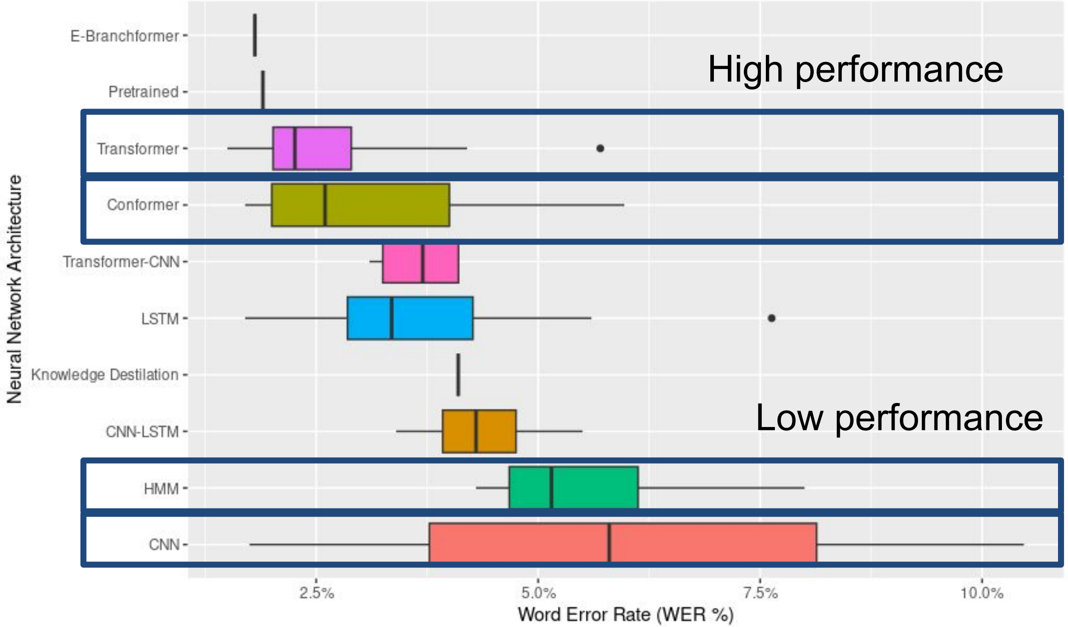


Figura 2. Distribution of Word Error Rate (WER%) in different speech recognition model architectures.

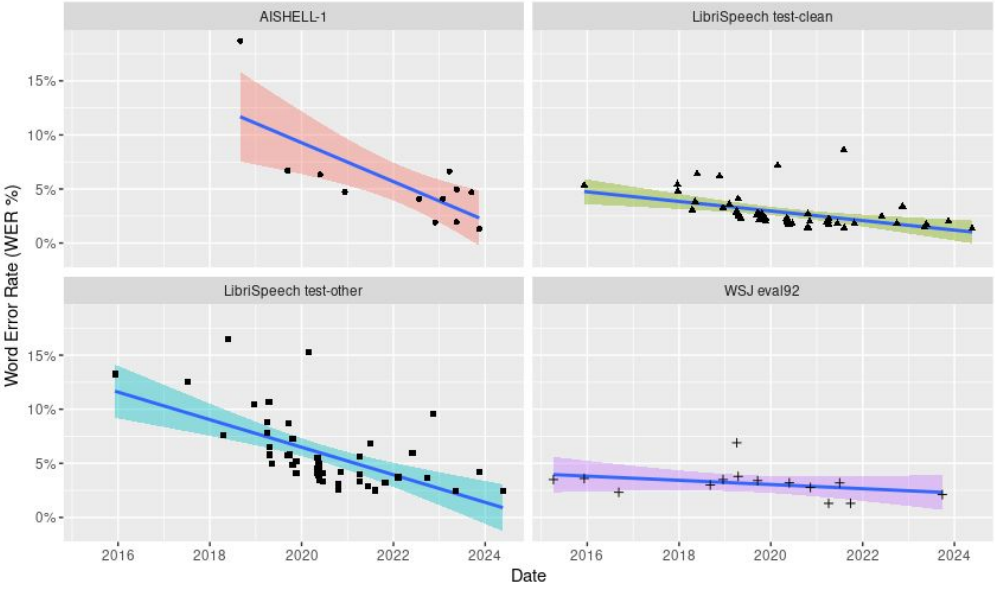
RESULTS

Trend fits of the form

$$y=a + bx$$

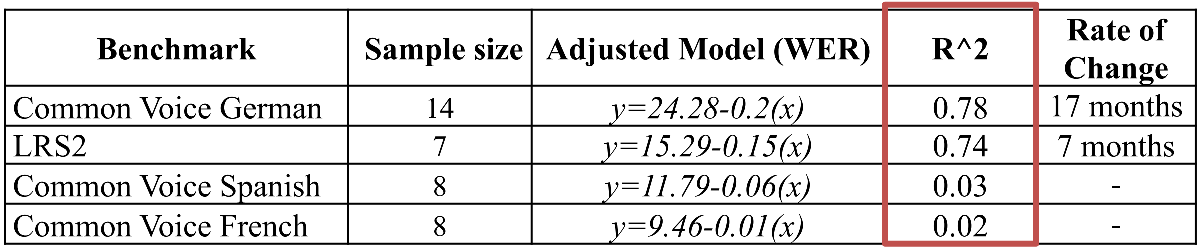
Ec 2. where x is the explanatory variable, y is the dependent variable, b is the slope of the regression line and a is the intercept (the value of y when x=0). For an in-depth consultation of the method, it is suggested to consult Su et al., (2012).

Benchmark	Sample size	Adjusted Model (WER)	R^2	Rate of Change
AISHELL-1	14	*y=17.76-0.15(x)*	0.52	7 months
LibriSpeech test-other	48	*y=12.5-0.10(x)*	0.35	10 months
LibriSpeech test-clean	57	*y=17.76-0.15(x)*	0.2	7 months
WSJ eval 92	15	*y=3.96-0.01(x)*	0.11	-

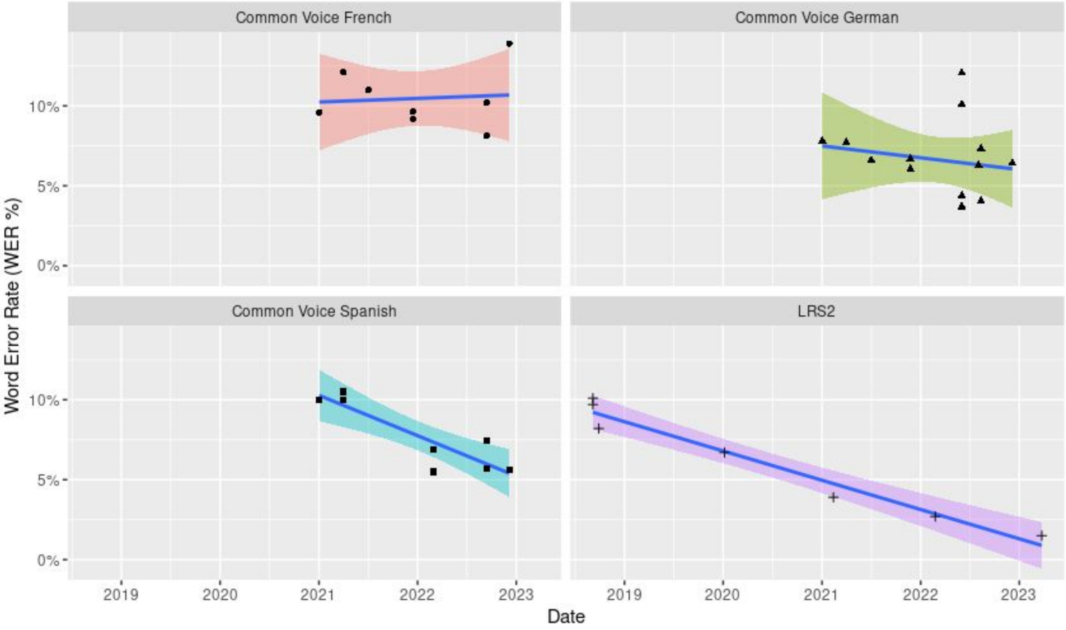


RESULTS

The estimates presented in these results exhibit **high uncertainty due to the small sample size.**



Benchmark	Sample size	Adjusted Model (WER)	R^2	Rate of Change
Common Voice German	14	*y=24.28-0.2(x)*	0.78	17 months
LRS2	7	*y=15.29-0.15(x)*	0.74	7 months
Common Voice Spanish	8	*y=11.79-0.06(x)*	0.03	-
Common Voice French	8	*y=9.46-0.01(x)*	0.02	-



CONCLUSIONS:

- The **architectures** with the **lowest error rate** were identified as **Transformer, Conformer and E-Branch Former.**
- The **models** evaluated in the **LibriSpeech Test Clean benchmark** present the **lowest error rate (WER).**
- Unfortunately, a **high uncertainty in the estimation of trends** in the speech recognition models stands out.
  - The trend fits for the analyzed benchmarks yielded **R^2** values **lower** than **0.78**, indicating an **insufficient fit** of the models to the data.

CONCLUSIONS:

Please Report Your Compute

CONCLUSIONS:

- The development of this project and participation in the "**Carreras con Impacto" program** have provided me with **valuable tools to increase** my chances of **success** in my future in science and **realign my career goals towards greater global impact.**

REFERENCES

- Dan Hendrycks. Introduction to AI Safety, Ethics and Society. Taylor & Francis, (forthcoming). ISBN: 9781032798028. URL: www.aisafetybook.com
- Jaime Sevilla, Anson Ho, and Tamay Besiroglu. ‘Please Report Your Compute’. Commun. ACM 66, no. 5 (May 2023): 30–32. https://doi.org/10.1145/3563035.
- NithyaKalyani, A., & Jothilakshmi, S. (2019). Speech summarization for tamil language. In Intelligent Speech Signal Processing (pp. 113-138). Academic Press.
- Su, X., Yan, X., & Tsai, C. L. (2012). Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics, 4(3), 275-294.

AI Safety

September 2024

Analysis of Progress in Speech Recognition Models

Miguel Angel Peñaloza

penaloza@cicese.edu.mx

CARRERAS CON
IMPACTO