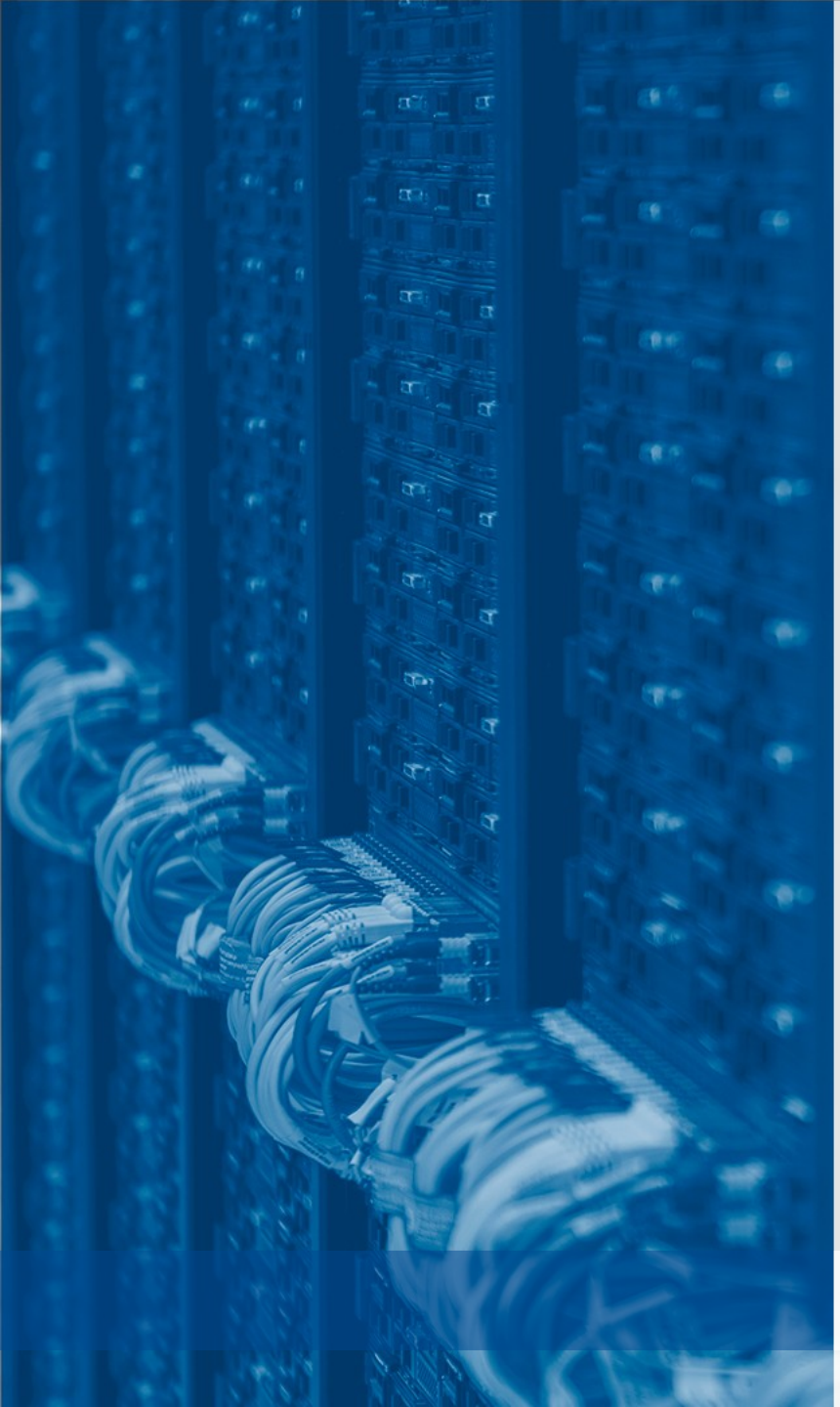


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Automatic Validation of the Non-Validated Spanish Speech Data of Common Voice 17.0

Carlos Daniel Hernández Mena

Outline

- Mozilla Common Voice
- Validation Methodology
- Results
- Deliverables
- Discussion
- Conclusions and Further Work

Mozilla Common Voice

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Validated Data in Different Languages

Español

Hours: 2274
Hablantes: 26437
Validation Progress: 20%
Frases: 1077425

CONTRIBUTE

Inglés

Hours: 3680
Hablantes: 95965
Validation Progress: 72%
Frases: 1601650

CONTRIBUTE

Catalán

Hours: 3373
Hablantes: 36638
Validation Progress: 85%
Frases: 1289896

CONTRIBUTE

Italiano

Hours: 443
Hablantes: 7272
Validation Progress: 82%
Frases: 919722

CONTRIBUTE

Portugués

Hours: 222
Hablantes: 3657
Validation Progress: 80%
Frases: 43394

CONTRIBUTE

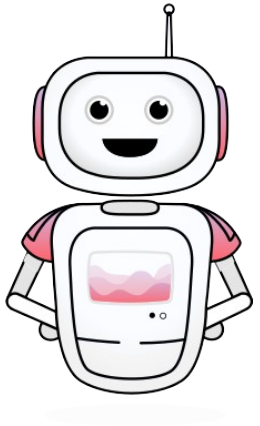
Francés

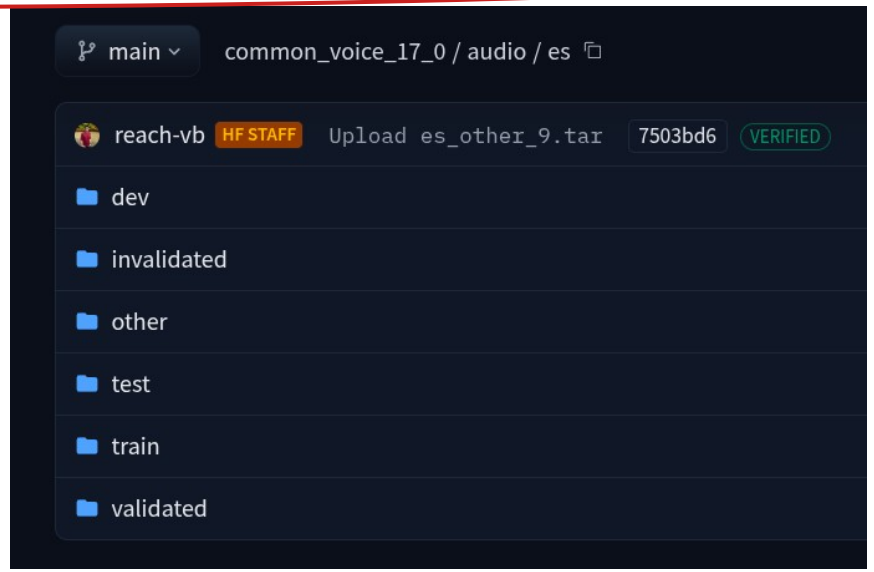
Hours: 1173
Hablantes: 20053
Validation Progress: 90%
Frases: 1604032

CONTRIBUTE

Common Voice 17.0 (Spanish)

Versión	Fecha	Tamaño	Horas grabadas	Horas validadas	Licencia	Número de voces	Formato de audio
✓ Common Voice Delta Segment 20.0	11/12/2024	142,46 MB	8	4	CC-0	140	MP3
Common Voice Corpus 20.0	11/12/2024	47,45 GB	2244	579	CC-0	26.290	MP3
Common Voice Delta Segment 19.0	18/9/2024	152,66 MB	8	4	CC-0	43	MP3
Common Voice Corpus 19.0	18/9/2024	47,31 GB	2237	576	CC-0	26.150	MP3
Common Voice Delta Segment 18.0	19/6/2024	195,72 MB	10	10	CC-0	65	MP3
Common Voice Corpus 18.0	19/6/2024	47,17 GB	2230	572	CC-0	26.107	MP3
Common Voice Delta Segment 17.0	20/3/2024	445,95 MB	16	13	CC-0	347	MP3
Common Voice Corpus 17.0	20/3/2024	46,97 GB	2220	562	CC-0	26.042	MP3

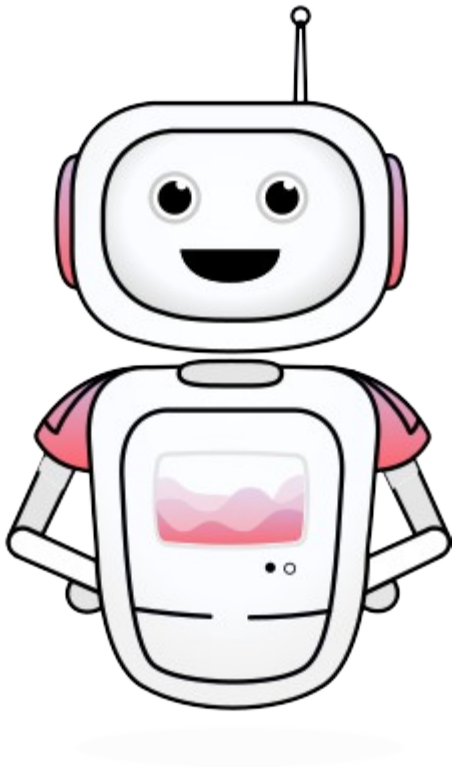




Barcelona Supercomputing Center
Centro Nacional de Supercomputación

We worked with the version in HF!

Mozilla Common Voice



Categories

Validated (**Train**, **Test** and **Validation** are here): At least two more positive votes than negative ones.

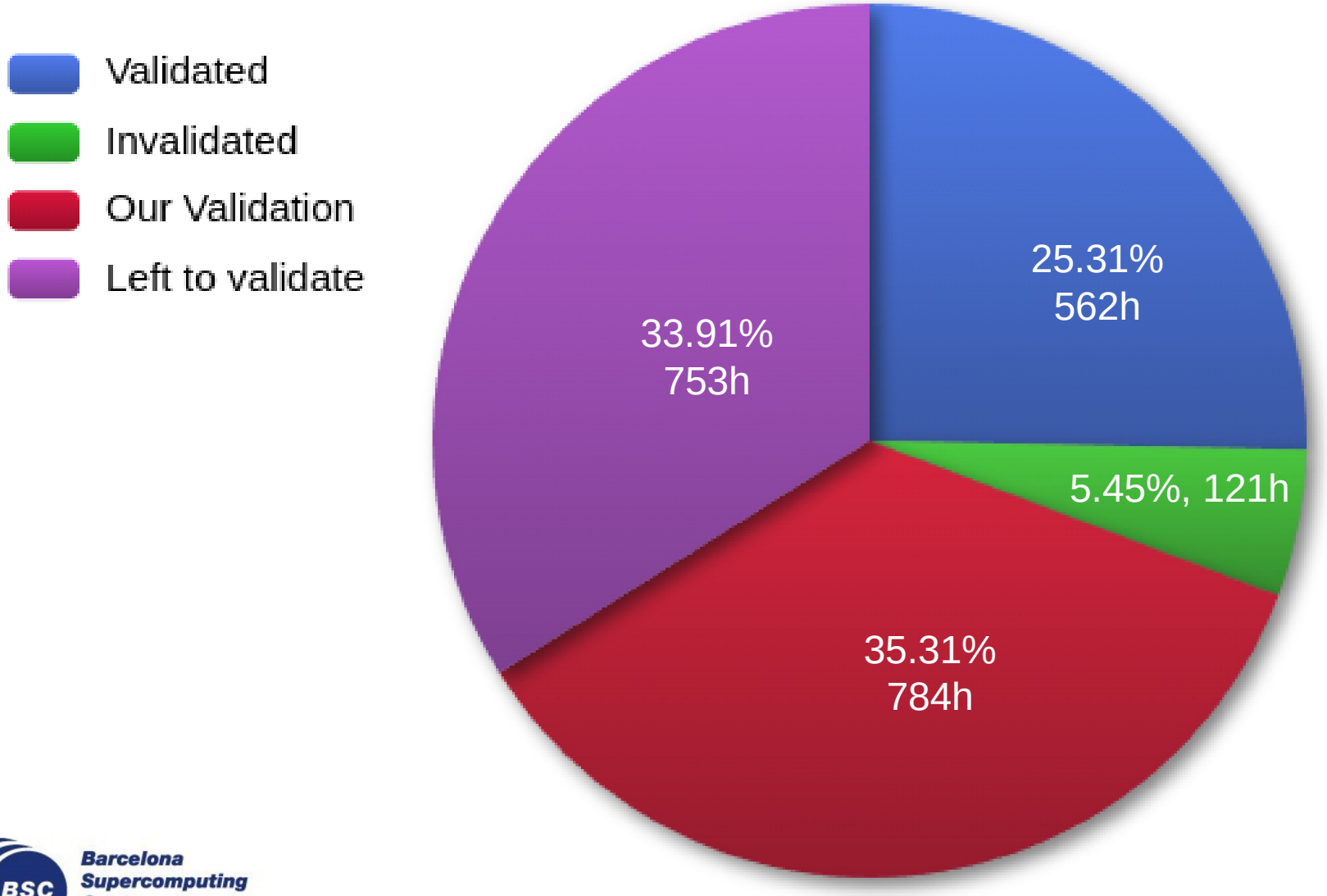
Invalidated: At least two more negative votes than positive ones.

Reported: inappropriate content or other issues.

Other: Not sufficient votes to make a decision

In this work we focus in the category “other”.

Portions of Common Voice 17.0



In original CV17 around 1500 hours (~70%) of audio are in the split called "**other**".

The portion validated by us is the largest !!!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Validation Methodology

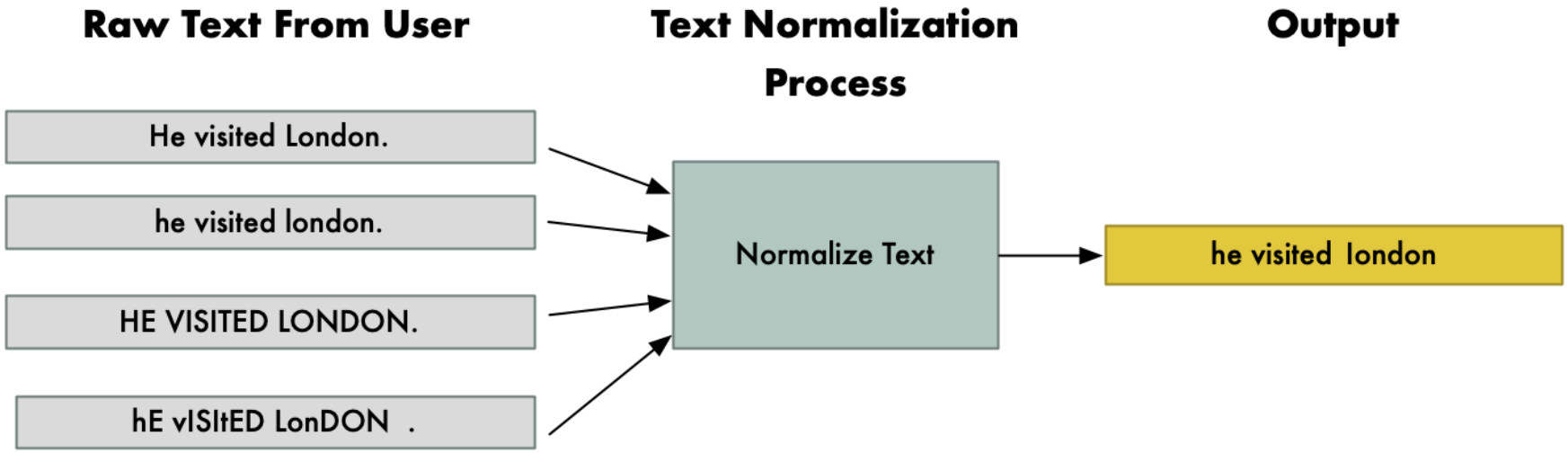
Whisper Out-of-the-Box

Direct validation = Perfect Matches

OpenAI Whisper

We seek for “Perfect Matches” between the Reference and Whisper’s Transcriptions!

Normalization of the Transcripts

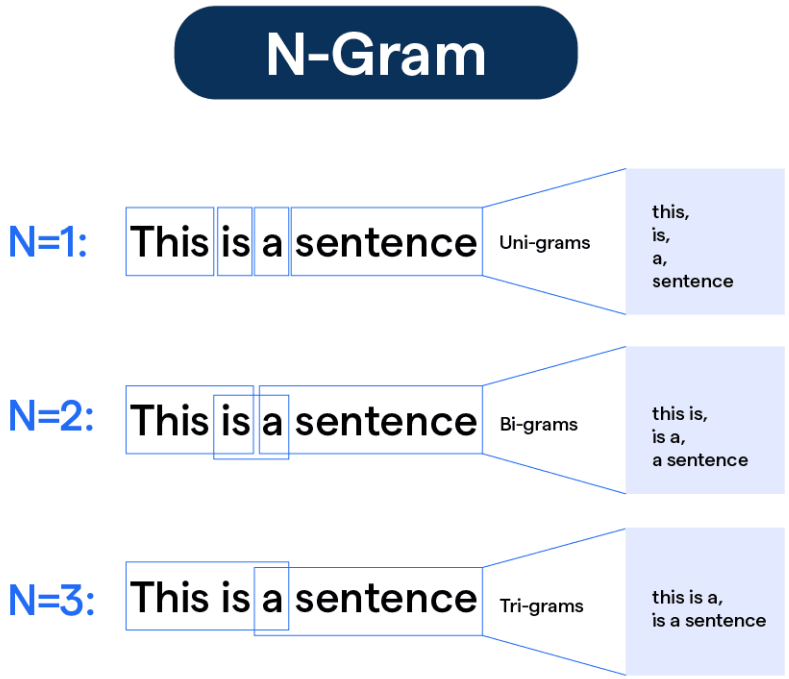


We applied normalization to both the reference and Whisper transcriptions.

- Lower case
- Remove punctuation
- Removing characters not belonging to Spanish alphabet (ä, ë, ô, ö ,etc.)

Classic ASR systems do not accept punctuation!

We didn’t use a Language Model (n-gram)



In large vocabulary ASR systems, there are many possible word choices. N-grams help to narrow down the search space, making the process more efficient. Instead of searching through every possible word, the system can focus on the most probable sequences.

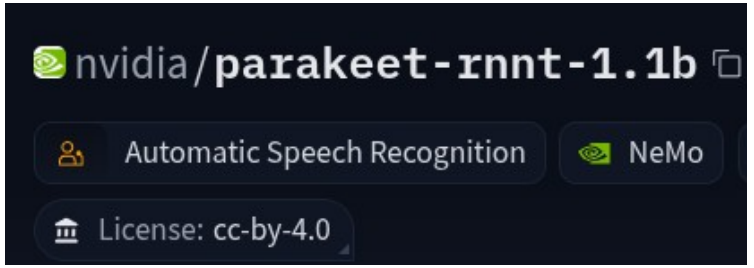
- <header - information ignored by applications>
- \data\
- ngram 1=9
- ngram 2=11
- ngram 3=3
- \1-grams:
- -0.8953 <unk> -0.7373
- -0.7404 </s> -0.6515
- -0.7861 <s> -0.1764
- -1.0414 When -0.4754
- -1.0414 will -0.1315
- -0.9622 the 0.0080
- -1.4393 Stock -0.3100
- -1.0414 Go -0.3852
- -0.9622 Up -0.1286
- \2-grams:
- -0.3626 <s> When -0.1736
- -1.2765 <s> the 0.0000
- -1.2765 <s> Up 0.0000
- -0.2359 When will 0.1011
- -1.0212 will </s> 0.0000
- -0.4191 will the 0.0000
- -1.1004 the </s> 0.0000
- -1.1004 the Go 0.0000
- -0.6232 Stock Go 0.0000
- -0.2359 Go Up 0.0587
- -0.4983 Up </s>
- \3-grams:
- -0.4260 <s> When will
- -0.6601 When will the
- -0.6601 Go Up </s>
- \end\

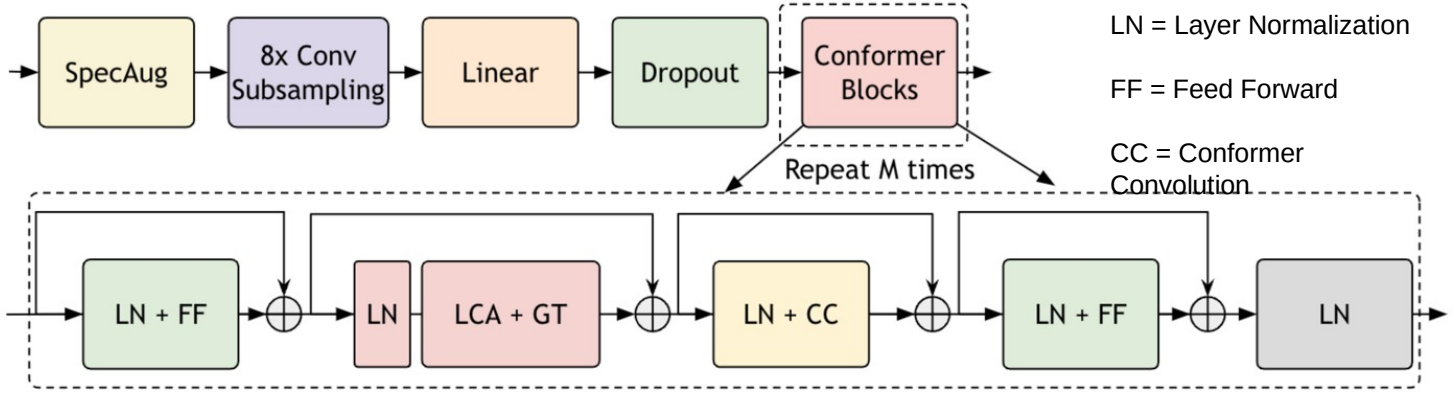
The ASR system will only choose words present in the n-gram model. We don’t want that!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

NVIDIA’s Parakeet Architecture

Indirect validation = Training an ASR model with our validated data





LN = Layer Normalization

FF = Feed Forward

CC = Conformer Convolution

RNNT = Recurrent Neural Network Transducer

Architecture of the NVIDIA Parakeet encoder with blocks of downsampling and subsampling, conformer encoder blocks with limited context attention (LCA), and global token (GT).

Bad data can’t produce Good models!

As far as we know, there is no official model of Parakeet in Spanish trained by NVIDIA!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

https://developer.nvidia.com/blog/pushing-the-boundaries-of-speech-recognition-with-nemo-parakeet-asr-models/

Hugging Face Leaderboard

Leaderboard

Speech Recognition

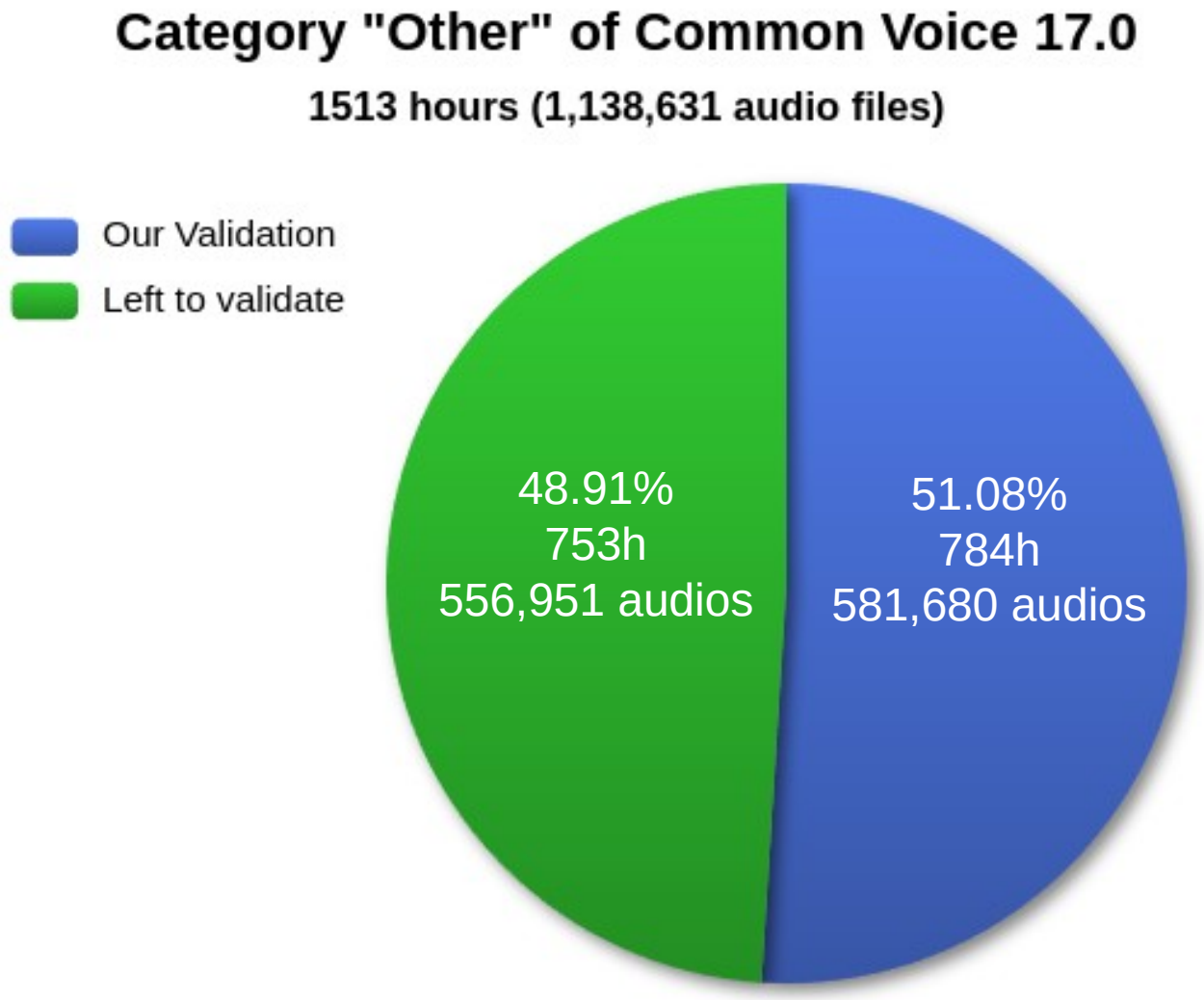
model	Average WER ⬇️	RTFx ⬆️	AMI	Earnings22	Gigaspeech	LS Clean	LS Other	SPGISpeech	Tedlium
nvidia/canary-1b	6.5	235.34	13.9	12.19	10.12	1.48	2.93	2.06	3.56
nyrahealth/CrisperWhisper	6.67	84.05	8.71	12.89	10.24	1.82	4	2.7	3.2
nvidia/parakeet-tdt-1.1b	7.01	2390.61	15.87	14.49	9.52	1.4	2.6	3.16	3.59
nvidia/parakeet-rnnt-1.1b	7.12	2053.15	17.01	13.94	9.89	1.45	2.5	2.93	3.83
nvidia/parakeet-ctc-1.1b	7.4	2728.52	15.67	13.75	10.28	1.83	3.51	4.02	3.57
openai/whisper-large-v3	7.44	145.51	15.95	11.29	10.02	2.01	3.91	2.94	3.86
nvidia/parakeet-tdt_ctc-110m	7.49	5345.14	15.89	12.37	10.52	2.4	5.22	2.54	4.07
nvidia/parakeet-rnnt-0.6b	7.5	2815.72	17.4	14.66	10.01	1.62	3.02	3.32	3.85
distil-whisper/distil-large-v3	7.52	214.42	15.16	11.79	10.08	2.54	5.19	3.27	3.86
nvidia/parakeet-ctc-0.6b	7.69	4281.53	16.46	14.26	10.39	1.88	3.8	3.89	3.77
openai/whisper-large-v3-turbo	7.83	200.19	16.13	11.63	10.14	2.1	4.24	2.97	3.57
openai/whisper-large-v2	7.83	144.45	16.74	12.05	10.67	2.83	5.14	3.87	3.9

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

https://huggingface.co/spaces/hf-audio/open_asr_leaderboard

Results

Validation in Numbers



We did not find a “perfect match” in the “left to validate” ones.

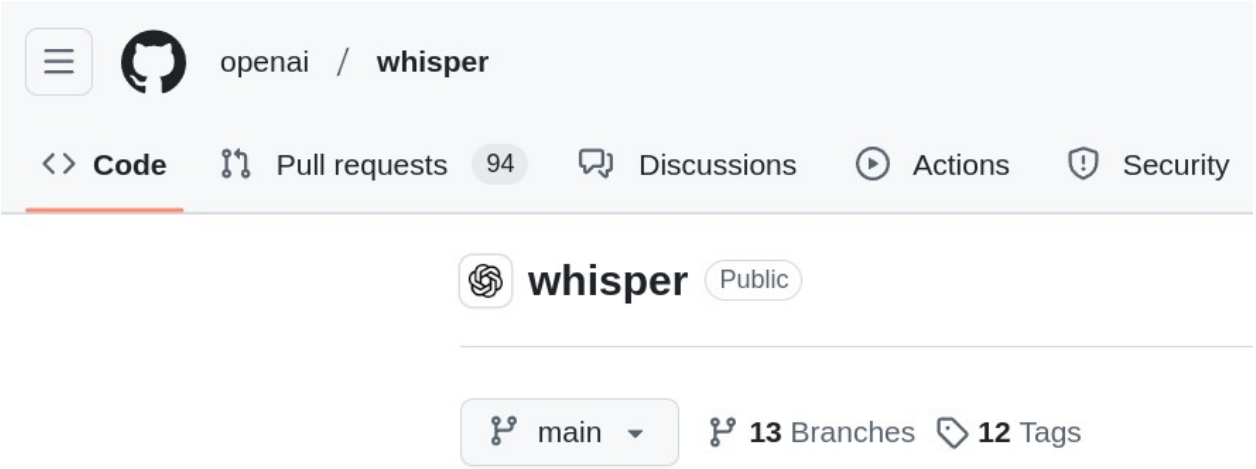
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

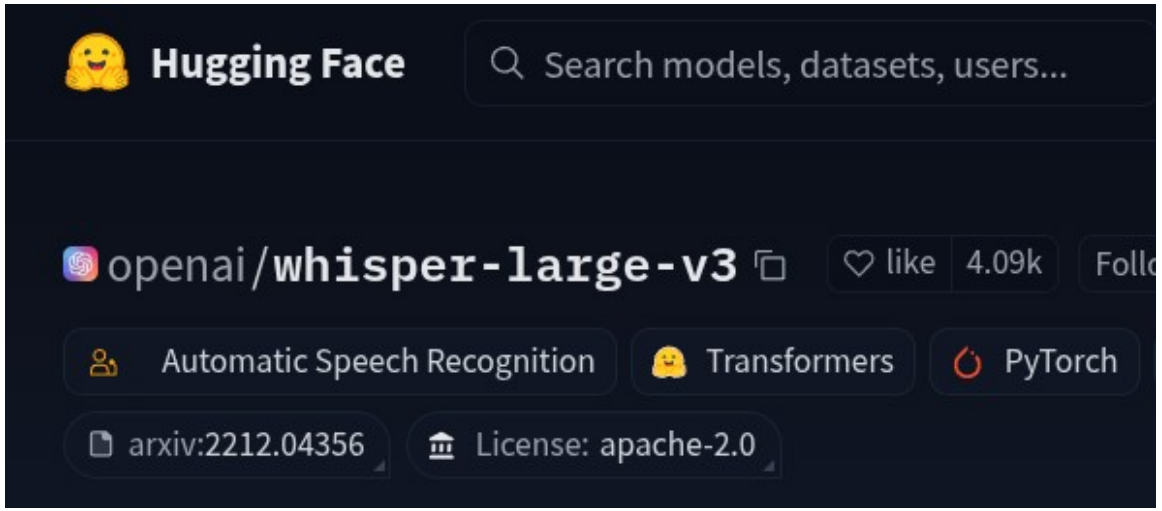
WER and CER Results

Model	Split	WER (%)	CER (%)
CV17 Validated	Test	5.13	1.69
	Dev	4.66	1.41
CV17 Other	Test	5.23	1.80
	Dev	4.85	1.53
CV17 Combined	Test	**3.93**	**1.29**
	Dev	**3.55**	**1.05**
OpenAI Whisper large	Test	4.97	1.81
	Dev	4.21	1.45
Whisper-large-v3	Test	5.15	1.84
	Dev	4.34	1.48

Table 1: Performance of the models trained with distinct subsets of Common Voice compared to the performance of two different versions of Whisper.

OpenAI Whisper vs Hugging Face Whisper





We have observed that they don’t provide the same results.

Deliverables

Deliverables in Hugging Face

projecte-aina/**parakeet-rnnt-1.1b_cv17_es_ep18_1270h**

NeMo | License: apache-2.0

Split CV17	WER (%)	CER (%)
Test	3.93	1.29
Dev	3.55	1.05

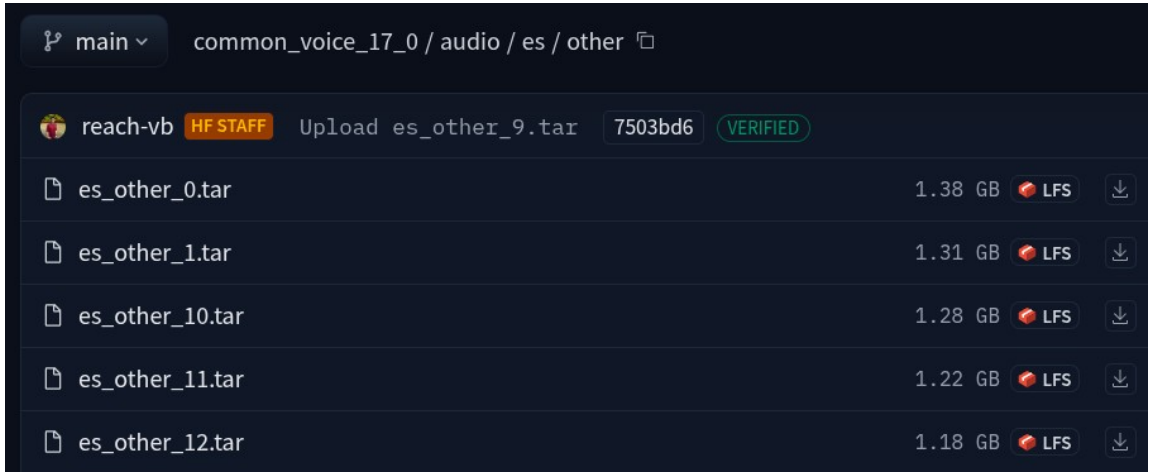
Hugging Face | Search models, datasets, users...

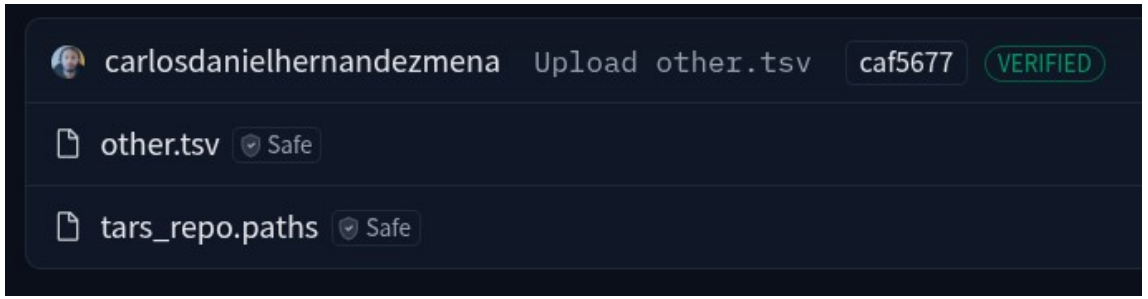
Datasets: projecte-aina/**cv17_es_other_automatically_verified**

Tasks: Automatic Speech Recognition | Languages: Spanish | Size: 100K<n<1M | Tags: common voice

784 hours and 50 minutes!

Our HF Repo Doesn’t Contain Audios





Our Repo takes the audio files from the Original Common Voice HF Repo!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

You will have to agree to the terms and conditions shown on the dataset card of Mozilla's HF Repo!

Discussion

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Possible Criticisms

The use of just one ASR system.

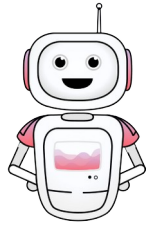
The use of one ASR system does not invalidate the results of other ASR systems.

The use of normalized transcriptions

he visited london

Normalized transcriptions enable compatibility with a broader spectrum of ASR systems.

It is very likely that Whisper was trained with Common Voice



Even with that, our Parakeet model outperformed Whisper!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Conclusions and Further Work

Conclusions

Our contributions to the community are:

A Parakeet model that beats Whisper in the Dev and Test of Common Voice 17.0

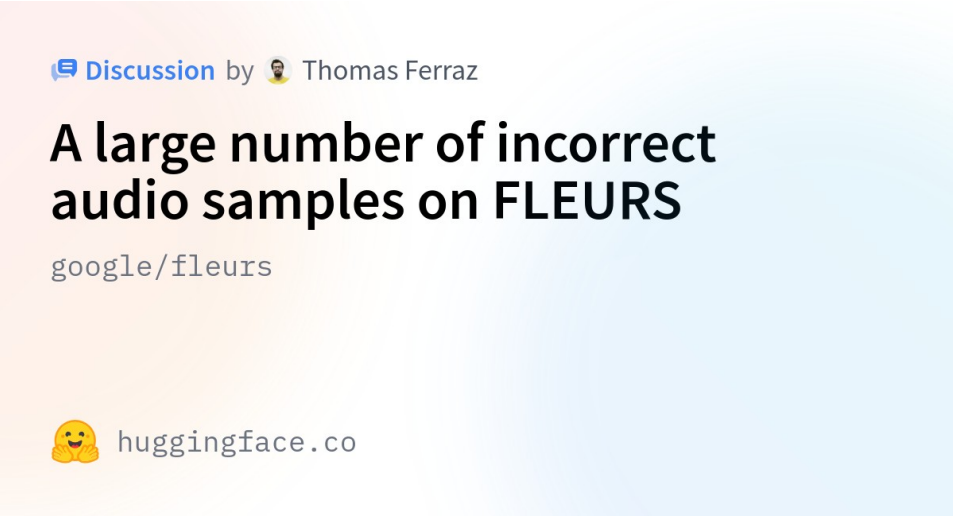
https://huggingface.co/datasets/projecte-aina/cv17_es_other_automatically_verified

An automatically validated corpus of 784 hours and 50 minutes.
https://huggingface.co/projecte-aina/parakeet-rnnt-1.1b_cv17_es_ep18_1270h

This contributions are publicly available in Hugging Face!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Further Work



We can apply this method to validated other datasets in the future!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Acknowledgements

ILENIA

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project **ILENIA** with reference **2022/TL22/00215337**.

Financiado por la Unión Europea
NextGenerationEU

GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

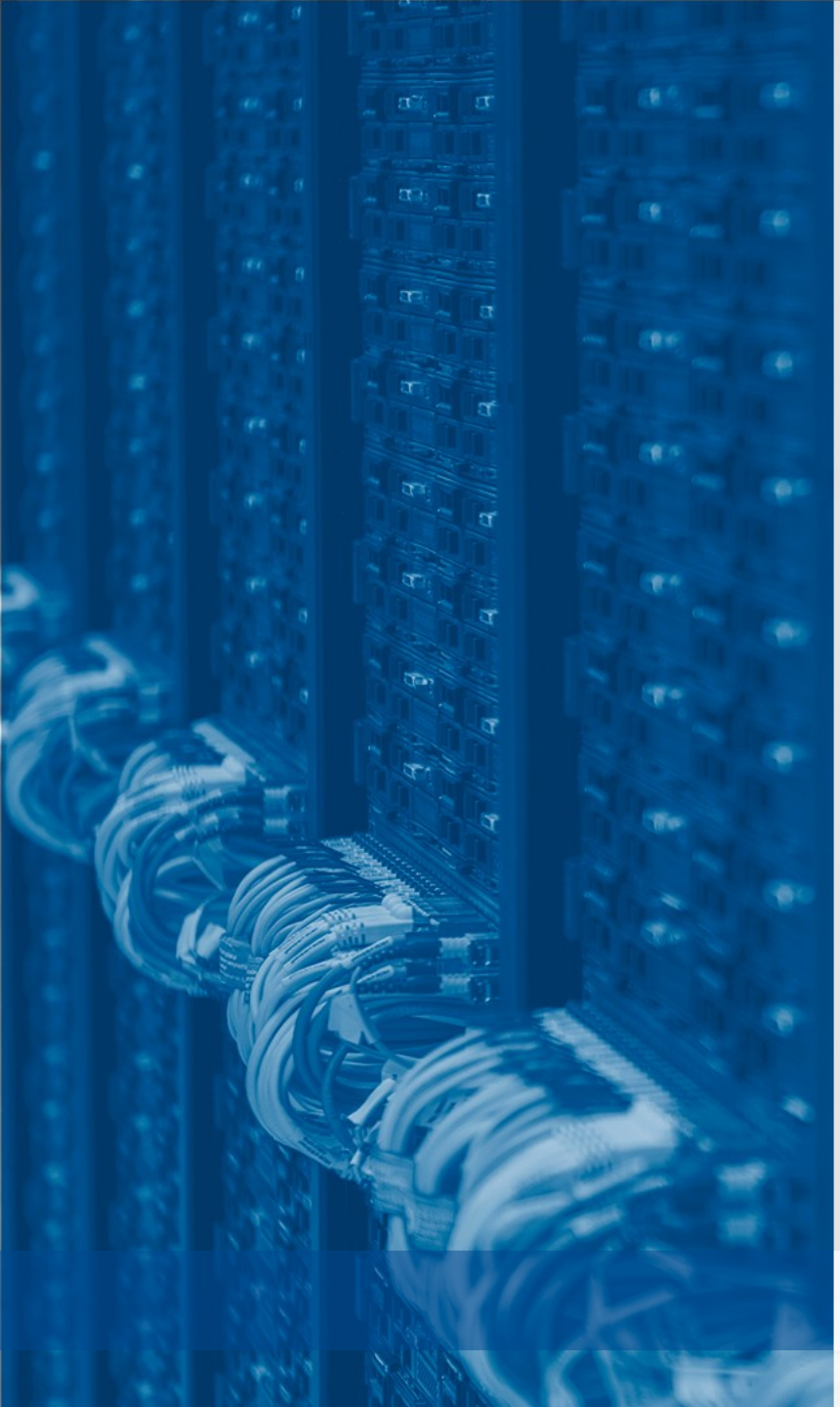
red.es

Plan de Recuperación, Transformación y Resiliencia

PERTE Nueva Economía De la Lengua

Questions?

Barcelona Supercomputing Center
Centro Nacional de Supercomputación



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Automatic Validation of the Non-Validated Spanish Speech Data of Common Voice 17.0

Carlos Daniel Hernández Mena