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

Carlos Daniel Hernández Mena[a], **Barbara Scalvini**[b], **Dávid í Lág**[b]
[a]Barcelona Supercomputing Center
carlos.hernandez@bsc.es
[b]University of the Faroe Islands
{barbaras,davidl}@setur.fo

Abstract

Mozilla Common Voice is a crowdsourced project that aims to create a public, multilingual dataset of voice recordings for training speech recognition models. In Common Voice, anyone can contribute by donating or validating recordings in various languages. However, despite the availability of many recordings in certain languages, a significant percentage remains unvalidated by users. This is the case for Spanish, where in version 17.0 of Common Voice, 75% of the 2,220 hours of recordings are unvalidated. In this work, we used the Whisper recognizer to automatically validate approximately 784 hours of recordings which are more than the 562 hours validated by users. To verify the accuracy of the validation, we developed a speech recognition model based on a version of NVIDIA-NeMo's Parakeet, which does not have an official Spanish version. Our final model achieved a WER of less than 4% on the test and validation splits of Common Voice 17.0. Both the model and the speech corpus are publicly available on Hugging Face.

1 Introduction

Developing Automatic Speech Recognition (ASR) systems requires extensive labeled speech data, which is costly and time-consuming to annotate manually. Crowdsourcing and automated methods help address these challenges by enabling efficient and consistent validation of data quality. For instance, Hernandez et al. (2018) used the Kaldi toolkit (Povey et al., 2011) to prepare the TED-LIUM corpus, while Krizaj et al. (2022) introduced a toolkit for automatic validation based on criteria like audio quality and transcription accuracy. Automated pipelines have also been applied in domain-specific scenarios (Romanovskyi et al., 2021), audio-visual speech recognition (Ma et al., 2023), and multilingual ASR systems from parliamentary archives (Nouza and Safarik, 2017; Kulebi et al., 2022; Helgadóttir et al., 2017). In this work, we employ OpenAI's Whisper ASR model to automatically validate the data by comparing the automatic transcription to a reference, which may or may not accurately reflect the content of the speech recording.

1.1 Objectives

Mozilla Common Voice (Ardila et al., 2019) is a multilingual, crowdsourced dataset of voice recordings that are validated based on user votes. Recordings are labeled as "validated" if they receive at least two more positive votes than negative ones. Recordings rejected by the community are labeled as "invalidated," while those with inconclusive results are categorized as "other."

We focus on the "other" category in Common Voice 17.0 (CV17), using Whisper-based ASR to validate recordings by matching transcriptions with references, as done in similar efforts for Icelandic data (Hernández Mena et al., 2024).

We evaluated our method with NVIDIA's Parakeet architecture (Galvez et al., 2024), comparing models fine-tuned on original CV17 validated data ($\sim$500 hours) and our validated data ($\sim$784 hours), both achieving a Word Error Rate (WER) $< 5\%$. Combining both subsets ($\sim$1284 hours) further reduced WER. Our validated dataset and best-performing model are publicly available on Hugging Face.

1.2 Paper Organization

This paper is organized as follows: Section 2 presents the final version of *The Corpus* shared in this work. Section 3 details the *Validation Methodology* used to automatically verify the

Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025), pages 58–63

CV17 speech recordings, which later became part of the corpus. In Section 4, we describe the development and fine-tuning of the *Acoustic Models* used to assess the effectiveness of the validation methodology. Finally, Section 5 concludes the paper with a summary of our contributions and suggestions for future work.

2 The Corpus

The corpus "Spanish Common Voice V17.0 Split Other Automatically Verified,"[1] as its name suggests, is the result of the automatic validation of the split called "other," which is part of the Spanish version of the Common Voice 17.0 corpus (CV17 for short). The corpus contains 784 hours and 50 minutes of audio across 581,680 recordings, surpassing the size of the "validated" category in CV17, which contains only 562 hours. Of these, 53 hours are allocated to the "test" and "dev" splits, while the remaining 509 hours belong to the "train" split. In comparison to the original CV17, our corpus contains only a single split called "other."

2.1 Audio Format

The audio files are distributed in the same format as the original CV17, with a sample rate of 48 kHz, a single channel, a bitrate of 64 kbps, and the MPEG-1 Layer 3 (MP3) codec.

2.2 Data Loader

In general, Hugging Face allows users to share datasets with others through dataset cards. A dataset card is a web page that contains the profile of the dataset. On this "web page," users can typically find documentation for the dataset, speech files, transcriptions, and metadata, as is the case with our corpus. Since the repository chosen to share our corpus is Hugging Face, the implications are that 1) the corpus has its own dataset card and, 2) the speech data can be accessed through the "datasets" Python library (Lhoest et al., 2021). In datasets, the object responsible for downloading the data from the dataset card, loading it into memory, and allowing Python to iterate over each recording in a for-loop is a "data loader." The data loader communicates with code executed by the Hugging Face website via the dataset card of the corpus. We programmed our data loader to download the data directly from the original CV17 repository, which means that our dataset card does not contain any audio files. The only information provided is a TSV file containing metadata for each recording in the corpus. Consequently, before downloading our corpus through the datasets library, it is important to agree to the terms and conditions shown on the dataset card for Mozilla Common Voice.[2]

2.3 Corpus Metadata

As explained in Section 2.2, the corpus metadata is contained in a TSV file that is stored in the dataset card. The information in the TSV was taken from the original CV17 with no changes; however, the rows in the TSV correspond only to the speech files validated by us. The columns of this TSV file are as follows: The `client_id` is a hexadecimal ID identifying the client (voice) that made the recording. The `path` field specifies the ID of the audio file followed by the extension ".mp3". The `sentence_id` is a hexadecimal ID of the speaker. The `sentence` field contains the sentence the user was prompted to speak. The `sentence_domain` indicates the context or scope to which the sentence belongs (this field is empty in all cases). The `up_votes` and `down_votes` fields represent the number of upvotes and downvotes, respectively, received by the audio file from reviewers. The `age` field denotes the age group of the speaker (e.g., teens, twenties, fifties), while the `gender` field specifies the speaker's gender (`male_masculine` or `female_feminine`). The `accents` field lists the speaker's accent(s) (e.g., España, México, Caribe, América Central), and the `variant` field refers to specific types of accents or pronunciation patterns associated with the speaker (this field is empty in all cases). The `locale` field indicates the locale of the speaker (the value is "es" in all cases). Finally, the `segment` field can either be empty or have the value "Benchmark."

3 Validation Methodology

Whisper (Radford et al., 2023) is one of the state-of-the-art multilingual speech recognition and translation models, available under the MIT license. It utilizes a Transformer-based architec-

[1] `https://huggingface.co/datasets/projecte-aina/cv17_es_other_automatically_verified`

[2] `https://huggingface.co/datasets/mozilla-foundation/common_voice_17_0`

ture with an encoder-decoder structure, trained via weak supervision on a massive dataset of multilingual speech (680 000 hours). The Whisper architecture supports both transcription and translation tasks.

As part of our validation methodology, we use OpenAI's Whisper to transcribe the speech recordings in the "other" category of CV17. If Whisper produces the same transcription as the reference, the recording is considered validated and added to the final corpus; otherwise, the recording is rejected. A total of 1,138,631 recordings were transcribed through this process, of which 581,680 (784 hours and 50 minutes) matched the reference transcription. This subset represents the total size of the corpus shared in this work. However, the reference transcriptions used in this process are not the original ones found in CV17 but have been normalized.

3.1 Normalization of the Transcripts

Reference transcriptions in CV17 include capitalization and punctuation. However, CV17 is used in experiments with a wide variety of ASR models and architectures, some of which do not accept punctuation marks as inputs. Additionally, we have detected that the Spanish portion of CV17 contains some characters not belonging to the Spanish alphabet (e.g., ä, ë, ô, ö). For this reason, the version of CV17 that we store is normalized as follows: 1) lowercase, 2) punctuation marks removed, and 3) letters not belonging to the Spanish alphabet are replaced with white spaces. In consequence, the same normalization is applied to the output transcriptions of Whisper during the validation process described in Section 3.

4 Acoustic Models

An indirect way to assess the correctness of our validation process is to evaluate how our validated recordings perform when training a real ASR model, as faulty data would hinder the production of a good acoustic model. For this purpose, we fine-tuned distinct models based on NVIDIA's Parakeet architecture, as described in Section 4.1.

It is important to note that, to the best of our knowledge, the official model (trained by NVIDIA) based on the specific Parakeet architecture we use in this work is only available in English; so, we chose this Parakeet architecture with the hope of making a meaningful contribution to the language technologies community.

4.1 NVIDIA's Parakeet

The Parakeet ASR models (Galvez et al., 2024), developed by NVIDIA as part of the NeMo framework (Kuchaiev et al., 2019), are state-of-the-art speech recognition systems offering high-accuracy English transcription. The Parakeet family includes four models: two with RNNT decoders and two with CTC decoders. This study employs the nvidia/parakeet-rnnt-1.1b model, ranked third on the Hugging Face speech recognition leaderboard.[3]

Built on the Fast Conformer architecture (Rekesh et al., 2023), an optimized version of the Conformer (Gulati et al., 2020), Parakeet features efficient downsampling, enhanced convolutional kernels, and local attention mechanisms. These improvements reduce memory use while enabling accurate transcription of audio segments up to 11 hours long (Koluguri et al., 2024).

4.2 Results

Table 1 shows WER and Character Error Rate (CER) results for models based on the "Parakeet RNNT 1.1B" architecture, evaluated on the "test" and "dev" splits of CV17. The "CV17 Validated" model was fine-tuned on user-validated CV17 data (∼500 hours), the "CV17 Other" model on data validated by our methodology (∼784 hours), and the "CV17 Combined" model on the combined dataset (∼1284 hours).

All models were fine-tuned for 48 hours using NVIDIA H100 GPUs. The "CV17 Validated" and "CV17 Other" models used 12 GPUs each, while the "CV17 Combined" model used 32 GPUs. Checkpoint 17, corresponding to epoch 18, was selected for all models to ensure comparability.

Additionally, Table 1 shows results using the first version of OpenAI's Whisper and the latest version Whisper-large-v3. As can be seen, the official Whisper models tend to outperform the models "CV17 Validated" and "CV17 Other"; however, our "CV17 Combined"[4] model outperforms all other models in the table, demonstrating the effectiveness of our validation method.

[3] https://huggingface.co/spaces/hf-audio/open_asr_leaderboard

[4] https://huggingface.co/projecte-aina/parakeet-rnnt-1.1b_cv17_es_ep18_1270h

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.

It is important to note that we distinguish between OpenAI's Whisper (sourced from GitHub) and Whisper-large-v3 (sourced from Hugging Face). In various experiments conducted for this and other studies, we have observed that the Whisper model available on GitHub and the Whisper model available on Hugging Face yield different results, even when they are the same size (tiny, base, small, etc.).

4.3 The Use of a Unique ASR System.

One potential criticism of this work is that we did not use multiple ASR systems for the validation process, as was done in the previously cited study by Hernández Mena et al. (2024). In that study, a recording was considered validated if at least one of their four ASR systems produced the same transcription as the reference, or a "perfect match" as they termed it. With this in mind, we can infer that involving additional ASR systems in our validation process would result in more validated recordings, though it would not invalidate those already verified by our single ASR system. Due to constraints in time and computational resources, we made the decision to use only one ASR system; however, we believe our results remain valid and valuable under the current experimental conditions.

4.4 The Use of Normalized Transcriptions

Another aspect worth discussing is the normalization of the transcripts. Given that the original CV17 references include punctuation and capitalization, and Whisper is capable of generating transcriptions with those same features, why not compare the transcripts without normalization? The answer lies partly in Section 3.1, where we explain that our laboratory experiment with a wide variety of ASR systems. Ultimately, we seek data that is compatible with both current and future experiments, adaptable to the latest technology as well as systems that have already proven reliable. In this regard, normalized transcriptions enable compatibility with a broader spectrum of ASR systems, many of which are not designed to handle punctuation, as is the case of Parakeet.

4.5 Performance of Acoustic Models

Results in Table 1 demonstrate the Whisper's impressive performance, as it outperforms two out of the three models we developed for this study. This reinforces the capability of Whisper to handle diverse datasets effectively, a result likely tied to the extensive training hours and resources invested in its development. Whisper's robustness in handling varied linguistic inputs, coupled with its high accuracy across CV17's "test" and "dev" splits, highlights its value as a benchmark model in automatic validation processes.

However, our best model, "CV17 Combined," achieves lower WER and CER than Whisper, suggesting that our validation method successfully curated a high-quality dataset for Spanish ASR. Although Whisper's performance is consistent with expectations given its extensive training set, and it was likely trained on a version of Mozilla Common Voice in Spanish that may introduce a bias enhancing its transcription accuracy on our test data, our results demonstrate that a carefully validated, language-specific corpus can yield models that not only compete closely with but even surpass larger-scale models.

These findings underscore the importance of targeted, language-specific model training, even in an era where large-scale, multilingual models dominate ASR.

5 Conclusions and Further Work

Crowdsourcing platforms are vital for ASR development, offering affordable and diverse data collection. However, manual validation limits their efficiency. This study demonstrated the potential of automatic validation for the Spanish subset of Common Voice 17.0 (CV17) using a Whisper-based ASR system. Our best Parakeet model

trained with the extended dataset, "CV17 Combined", outperformed both OpenAI's Whisper and Whisper-large-v3, showcasing the benefits of automated validation. Future work could explore applying this approach to other datasets (e.g., Voxforge[5]) and languages, especially low-resource ones, which could gain significantly from automated dataset expansion despite potential model performance challenges.

Acknowledgments

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.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Daniel Galvez, Vladimir Bataev, Hainan Xu, and Tim Kaldewey. 2024. Speed of light exact greedy decoding for rnn-t speech recognition models on gpu. *arXiv preprint arXiv:2406.03791*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Inga Rún Helgadóttir, Róbert Kjaran, Anna Björk Nikulásdóttir, and Jón Gudnason. 2017. Building an asr corpus using althingi's parliamentary speeches. In *Interspeech*, pages 2163–2167.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*, pages 198–208. Springer.

Carlos Daniel Hernández Mena, Thorsteinn Dadi Gunnarsson, and Jón Gudnason. 2024. Samrómur milljón: An asr corpus of one million verified read prompts in icelandic. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14305–14312.

Nithin Rao Koluguri, Samuel Kriman, Georgy Zelenfroind, Somshubra Majumdar, Dima Rekesh, Vahid Noroozi, Jagadeesh Balam, and Boris Ginsburg. 2024. Investigating end-to-end asr architectures for long form audio transcription. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13366–13370. IEEE.

Janez Krizaj, Jerneja Zganec Gros, and Simon Dobrisek. 2022. Validation of speech data for training automatic speech recognition systems. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1165–1169. IEEE.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al. 2019. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*.

Baybars Kulebi, Carme Armentano-Oller, Carlos Rodríguez-Penagos, and Marta Villegas. 2022. Parlamentparla: A speech corpus of catalan parliamentary sessions. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 125–130.

Quentin Lhoest, Albert Villanova Del Moral, Yacine Jernite, Abhishek Thakur, Patrick Von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. 2021. Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846*.

Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. 2023. Auto-avsr: Audio-visual speech recognition with automatic labels. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jan Nouza and Radek Safarik. 2017. Parliament archives used for automatic training of multi-lingual automatic speech recognition systems. In *Text, Speech, and Dialogue: 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings 20*, pages 174–182. Springer.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023.

[5] https://www.voxforge.org/

Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, et al. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

O Romanovskyi, I Iosifov, O Iosifova, V Sokolov, F Kipchuk, and I Sukaylo. 2021. Automated pipeline for training dataset creation from unlabeled audios for automatic speech recognition. In *International Conference on Computer Science, Engineering and Education Applications*, pages 25–36. Springer.