Analysis of the Evolution Path of the NVIDIA Granary Speech Dataset

Jiongzeng Ye[1*]

[1]School of Computer Science, Guangdong University of Science and Technology, Dongguan, China

Abstract. Today in the development of voice AI, it faces a big obstacle, called a “linguistic ecosystem imbalance” because the amount of minor language data is very scarce, which makes the model performance difficult, and making it unfair. Grannary: It is Nvidia’s open-sourced speech dataset project announced on August 2025 It has collected around 1 million hours of people’s voice audio. NVIDIA Granary is the first industrial scale speech dataset to cover many minor European languages. It will thus be a landmark for the task. This article seeks to comprehensively understand the development of Granary by researching and comparing versions of the granary: Scale up, language up, quality up, and up to ethics. In this paper, not only to fill the gap on the research about the evolutionary trend of a single dataset, but also to get a concrete industrial-level data building pattern; In the future, it can provide very valuable advice to how to build an even more inclusive, robust and trusted multilingual speech intelligence.

1 Introduction

In recent years with the fusion of voice data sets and artificial intelligence technologies, the voice recognition systems have been transferred from laboratory to every household and become the center of human computer intercommunication. From smart homes with voice commands to real-time translation during conferences in multinational corporations, the scope of voice technology applications is getting wider and wider. But there is actually big "language void" behind this prosperity scene of Voice AI. The excellent performances of mainstream ASR models are mainly based on the use of plenty of high quality data, which is dominated by mainstream language like English, Chinese and Spanish. This is in sharp contrast with the thousands of minority languages and dialects around the world, which are in a “data desert” for lack of data. Based on the European Language Diversity report, more than 60 known regional languages and minority languages exist in the EU. The smart ones that don’t have data aren’t really digital, and can’t be integrated into the modern smart ecosystem. This data imbalance is not just a problem about the data. This is also a social issue about the protection of different cultural groups and equal access to information.

Trends within this industry are definitely going to be towards larger and more comprehensive unified speech models that can speak many languages at once. An ordinary example might beguile. The USM has been a breakthrough in the field of artificial

* Corresponding author’s email: yjz3289562536@hotmail.com

intelligence automatic speech recognition, it can support more than 100 languages, tens of millions of hours of diverse speech training [1]. This is a new frontier in multilingual speech research and in that sense a new milestone, but it's also just this little nod that we put together a large and careful selection of multilingual data, it matters, it should matter, it should make progress. And the fact that USM works really well, it makes me think about the future of the AI world. The paper thinks the future of voice AI is going to be a foundation model - so you're training it on lots and lots of different language stuff, and there'll be lots different kinds of domains and different kinds of languages to use it on.

In this case, NVIDIA started and sustains a great amount multilingual voice data sets named "Granary". Granary is a large speech recognition and translation dataset for 25 European languages, created by NVIDIA and others. This dataset was made by using pseudo-labelling and it works for doing speech processing in language with few resources. The dataset is about 643,000 hours of speech data, which has gone through careful screening and processing. This can further help improve data quality and reduce mistakes in the data. Granary is a static data warehouse and also resembles a dynamic and continually improving ecosystem, with many iterations and revisions. it has laid down a good foundation for training the next generation of multilingual big models. But the current academic works most focus on how to use the granary to train a better model and finally obtain the results of the model However, there are very few studies that have followed and analyzed the growth of the Granary dataset itself-how it came from an initial idea through a series of important technical and engineering decisions to becoming a full-fledged dataset. The paper will then analyze the growth of the Granary Dataset. To comprehend the evolutionary path mentioned above is to guide designing better methods for constructing a data set in the future, which are more efficient, fair, and sustainable.

Although NVIDIA and its partners have not yet released an official version of the 'Granary' dataset (such as V1, V4, etc.), this study introduces a conceptual version framework to analyze the potential evolution of the dataset. The models of these four stages (V1-V4) do not represent officially documented versions, but are reconstructed based on publicly available materials (including NVIDIA's technical blogs and the 2025 Interspeech paper).By organizing these factual sources into a structured sequence, this framework helps illustrate how the 'Granary' dataset was developed through key stages such as data collection, pseudo-labelling, quality screening, governance, and public release.

2 Literature review

2.1 The main viewpoints of previous research

As large-scale datasets become the cornerstone of artificial intelligence development, the transparency and ethical considerations in their construction have increasingly become focal points of academic attention. The framework proposed by Gebru et al., "Datasheets for Datasets," has made a foundational contribution in this regard. The study advocates that datasets should provide standardized information—such as their motivation, composition, collection process, preprocessing steps, and potential biases—similar to the way physical products come with manuals, in order to promote traceability, auditability, and responsible use of datasets [2]. Framework provides major governance principles for dataset lifecycle management which will be referred to later, industry focus on data quality moves from technical metrics alone into ethics and society.

As for multilingual speech recognition, there's not only a 'big problem' of data desert at the Macro, but also a 'small problem' of data bias at the Micro. For resources-rich language like English and its training data can't cover all the internal diversity and lead to serious

fairness issues. The milestone study by Koenecke et al. has shown that the WER for commercial ASR systems recognizing African American English speech is almost double what it is for white speech, and the fundamental root cause of such systemic bias is due to training data being unbalanced with respect to different dialects/accents. Therefore, the rise of NVIDIA Granary and the strategy for language expansion is much more than just filling in the gap in minority languages [3].

It's worth mentioning that the research done by Akera et al. on the Kinyarwanda and Kikuyu languages is also an example of such evaluation [4]. This study also exposes the sensitivity of low-resource languages to the amount of training data and shows how tough the initial challenge is, which can be described as “Data Collection” problem Put it another way, it reveals the sparsity and foresight of Granary's industrialization and scale in data collection and annotation.

In recent years, the work on building multilingual speech datasets has attracted the attention of academia and industry. In order to fully understand the development of the NVIDIA Granary dataset, we need to study it within the scope of existing academic research. Beforehand, previous studies have given strong theoretical and analytical groundwork mainly focusing in four key aspects: dataset creation, technical difficulties, engineering tasks and fair ethical considerations. At the dataset construction methodology level, Nagrani et al. comprehensively expounded on the entire process of building an industrial-grade speech dataset. Their focus on the “trade-off between diversity and data quality” provided a straightforward framework of analysis for this paper [5]. This paper selects the multi-level evaluation indicators suggested by them for the quantitative analyses of the quality improvement level of Granary in each evolution stage. On the other hand, Ardila et al.'s models represent crowdsourcing where there are struggles in getting people on board and maintaining quality while trying to incorporate different languages [6]. Crowd's hybrids mixing industrial-level data collection with crowd sourcing and crowdsourced annotation: taking advantage of crowdsourcing benefit while overcoming inconsistency: Compare the pros and cons of these 2 models, it's instructive on how Granary makes its engineering choices.

In the next generation of self - supervised learning models, the value of large - scale multilingual data is completely unleashed. Conneau et al. showed that with pre-training on large-scale multilingual speech data, models can learn strong and transferable cross-lingual speech representations that benefit the recognition of low-resource languages [7]. That is to say, such datasets like granary possess some strategic value beyond what supervised learning would present to datasets, namely being raw material for a given recognition model to learn on; datasets are fuel and necessary for forming a universal speech foundation model.

To sum up, those earlier studies not only showed that granary was bound to emerge, but also supplied us with a set of means of examining its way of evolution in this study. The following text will give a detailed description of the development stages of Granary under above framework.

2.2 Relevant Theoretical Frameworks

Building on previous studies, this study mainly establishes an analytical model with the two theories as its foundation. Systematically analyze the evolutionary process of the NVIDIA granary dataset: the first is the data-centric Ai concept, the second is the dataset lifecycle management mode. The above two frameworks are respectively deep guidance on understanding the iterative logic of Granary from the perspective of philosophical methods and engineering management.

2.2.1 Data Centric AI Framework

This paradigm proposed by Andrew Ng indicates that in today's era where model architecture is already established, constantly and step by step improving data quality is more efficient for enhancing overall system performance compared to continuously innovating model architecture alone. The essential principle is to improve data quality by doing data cleaning, data improvement, and data consistency tagging, and improve model quality by doing data rebalancing. This framework gives top-level guidance and a guiding value for this study - Granary's evolution is not an unguided rise in function; each main update - from scale, to quality, to ethics -- is in fact a concentrated practice of the Data Centric concept. The focus of analysis should not be limited to what the dataset "has", but should also focus on what fundamental problems in model training it "solves" through data optimization. Therefore, the analysis of this study will focus closely on the core clue of how Granary drives potential model performance leaps through the dimension upgrade of data quality.

2.2.2 Dataset Lifecycle Management Model

This model originated from the field of data governance and was proposed and improved by Scheuerman et al. It regards datasets as dynamically evolving products, with a lifecycle that includes multiple stages such as planning, collection, cleaning, annotation, maintenance, distribution, and retirement. This framework provides a structured analytical tool for this study to microscopically examine the changes in Granary's center of gravity at different stages of development. As shown in Table 1, Granary's iterative process perfectly maps the rolling focus of each stage of the lifecycle: its early (v1-v2 stage) evolution focuses on the scale and language expansion of the "collection" stage, while in the middle and later stages (after v3 stage), it clearly shifts towards quality improvement and bias governance in the "cleaning", "labeling", and "maintenance" stages. Using this model, articles could be more than just a description of the dataset's static characteristics and be able to analyze the internal project's development logic and the prioritization of projects as a dynamic project.

Table 1. Development of Granary at Different Declaration Cycle Stages

Dataset lifecycle stages	GranaryV1-V2 (laying/expanding)	GranaryV3 (Quality)	GranaryV4 (Governance)
Planning/Collecting	Large scale collection of core, Nordic, and Baltic language systems	Targeted supplementation of scarce scenario data	Compliance and Ethical Source Planning
Clear/labeled	Basic quality filtering, text transcription	Advanced enhancement (noise), speaker metadata annotation	Bias auditing, data traceability
Maintenance/Distribution	Establish a basic distribution pipeline	Optimize format and support subset construction	Exploring blockchain traceability and efficient version management

Framework merging and application: This study integrates the above two frameworks and forms a complete analytical perspective (as shown in table 1):

• The idea of data centeredness sets up the main aim of the analysis, which is to see the evolutionary benefit that datasets have in improving models.

• The lifecycle management model provides a certain way of investigation, that is to say, breaking down the activities in every stage of the lifecycle to show the development of the dataset.

This fusion framework is good at carrying out Granary's evolution analysis; it allows us to make the point that Granary went from chasing scale (collection) to increasing diversity (collection) to ensuring quality (cleaning, tagging) to improving governance (maintaining) is a scientifically correct course of development that's aligned with the data-centric mindset and is well taken care of. Each phase represents a direct reaction to the problems found in the preceding phase of Granary's life cycle.

3 Evolution analysis of NVIDIA Granary dataset versions

The development of data, especially the NVIDIA Granary data set, is not achieved overnight, but rather a focused and continuous iterative process based on three axes, scale, quality, ethics [8]. On different versions the evolutional progression embodies how the industry has progressively developed its understanding of multilingual speech recognition problems. and by analyzing public information and technical reports this study reconstructs the following 4 important paths of evolution:(as shown in Table 2)

Table 2. Overview of NVIDIA Granary Dataset Version Evolution

Version stage	core objective	Key technical characteristics	Representative progress
V1: Foundation stage	Achieve a breakthrough from 0 to 1 and establish a basic scale	Large scale data pipeline architecture, basic quality filtering	Covering 5-10 core Western European languages; The total duration reaches several hundred thousand hours
V2: Expansion stage	Addressing the issue of 'data desert' and expanding language diversity	Strategic language planning, integration of multiple data sources, and advanced deduplication	Expand the number of languages to over 20; Systematically incorporating into the Baltic language system (Lithuanian, Latvian, etc.); Total scale exceeds one million hours
V3: Quality stage	Enhance the robustness and fairness of the model in the real world	Tool kit integration (noise injection, RIR simulation), refined metadata annotation	Add speaker attribute labels (age, gender) and environmental noise labels; Targeted enhancement of data performance in noisy scenes
V4: Governance Stage	Building a responsible and trustworthy data ecosystem	Data traceability technology, bias audit tools, efficient governance framework	Exploring the use of blockchain technology for data source tracing; Publish a dataset bias assessment report; Support the construction of subsets according to ethical requirements

3.1 V1 Foundation stage

The foundation stage, focusing on verifying the feasibility of the large scale multilingual dataset project, the main task is to build a high throughput automatic data collection and processing pipeline. Data composition wise, we should give more weight to the Western European languages with resource superiority, like English, German, and French, so as to swiftly build up a certain data scale. Data processing mainly uses basic filtering methods, based on the energy and duration of signals. Annotation content is limited to text transcription. Stage two implementation provided Granary with the engineering foundation of an industrial grade data set and addressed the largest issue of large scale collection and storage.

3.2 V2 Expansion stage

In the growth phase following the application of engineering collection, Granary faced the problem of shortage of minority language resources. This stage used a language expanding method by following the language lineage and regions significance, purposefully putting in Baltic and Nordic language such as Lithuanian, Estonian, Finno, etc. By integrating authorized commercial corpus and partner data, and applying near duplicate audio detection and channel standardization technology, the quality of newly added data is ensured while expanding the scale. Thus, Granary has established its unique position in multilingual speech datasets due to its systematic coverage of scarce languages.

3.3 V3 Quality stage

The quality phase marks a paradigm shift in Granary's construction focus from "quantity" to "quality", with its core feature being the refinement of deep embedding and annotation systems.

• In terms of data augmentation, noise injection and room impulse response simulation (RIR Simulation) techniques are widely used to synthesize a large amount of speech data that matches complex acoustic environments such as vehicles and industrial equipment, aiming to improve the real-world robustness of the trained model.

• In terms of annotation system, speaker metadata (such as age group, gender) and environmental labels for voice have been added in addition to transcribed text. This change brings richer supervision information for model training, but even more important is the establishment of a prerequisite for data bias auditing and fairness research, which is an example of Granary evolving towards responsible construction.

3.4 V4 Governance Stage

The governance phase mainly involves setting up data governance mechanisms and moral regulations, at this point Granary maturely becomes an infrastructure.

• At Technological levels, try doing a blockchain base data trace framework to make sure data sources and annotations can be checked;

• In terms of governance, gradually build bias evaluation and reduction systems, release bias reports, and offer instruments to support the development of ethical data subsets. This shows that Granary's growth is now entering a phase where technology capabilities and governance systems are collaborating to advance the cause of responsible data within artificial intelligence.

To sum up, Granary's version has developed along a clear path: from first creating some level of scale for construction, to gradually diversifying and improving mid-term until reaching a mature state of governance and ethics. Each stage of development aims to solve the main problems of its time, constantly feeding back the creation of stronger, fairer, and more trustworthy multilingual speech intelligence.

4 Discussion

Based on an analysis of the development of the NVIDIA Granary dataset, we found that its development has obvious stage characteristics. It is also the reflection of our industrial development. This shift reflects an adjustment of technology.

Granary's development also follows a strong track alongside the development pattern of the AI industry as a whole. In the early stage it pays attention to construction and scale, which

is directly related to the need for large amount of data in deep learning; mid-term change to quality means that the whole industry focuses on model practicality; recent emphasis on ethical governance reflects the society's expectation of responsible AI.

The merit of this evolutionary model is that it can be referred to. In other fields can also use this reference path for dataset construction to avoid re-innovation exploration. especially in fields like medical imaging, autonomous car driving, etc., which all need very high-quality data, this staged and focused development approach would be very important.

5 Limitations and Future Work

5.1 Challenge

It is necessary to point out that this study has certain restrictions. first, because it is impossible to find all the technical content, some analysis made on reasonable inference from the public information about the Granary dataset may be inaccurate.

Secondly, this study mainly uses qualitative analysis methods, which can show development trends and patterns, but there is no quantitative data support. For example, it is not possible to accurately measure the specific impact of each stage's improvement on the final model performance.

5.2 Future Work

By means of the results of the study and limitations, we think these are pretty good ideas:

Firstly, more in-depth quantitative research can be conducted. future researchers can try to grab the data from other different versions of GRANARY and get the real deal of improvement at every step by doing experimental work. This will give more precise instructions about dataset construction. Second, it is suggested to pay attention to the effect on the dataset due to new technology Especially when it comes to new technologies like blockchain and federated learning, it may change the way we collect, annotate, and govern data, which is worth researching.

At last, Researchers can also try to apply these results for other areas. For example, in fields such as medical imaging and bioinformatics, it is worth exploring whether similar development paths can be borrowed.

6 Conclusion

After researching the evolution of the NVIDIA Granary dataset, the paper has reached the following conclusion:

First of all, building high-quality datasets is a kind of gradualness with a stage of scale growth first and then quality improvement from scratch, and finally governance improvement. The latter process also shows the increasing awareness of industry as well.

Second, datasets also need to keep up with technological advancements. From Granary's experience, good datasets are not only large but also high-quality and ethical.

Finally, the evolutionary model proposed by this paper is universal, which also provides a reference for the construction of dataset in other fields. Especially because of the rapid development of AI, it is even more necessary to construct datasets in a systematic way.

In general, Granary's development supplies a good case study on how to develop a high-quality dataset methodically. It will be not only benefit to the academic study, but it will also be beneficial to industry practice guidance.

References

1. Y. Zhang et al., "Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages," arXiv preprint arXiv:2303.01037, (2023). Available: https://arxiv.org/abs/2303.01037
2. T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for Datasets," Commun. ACM, vol. 64, no. 12, pp. 86–92, (2021), doi: 10.1145/3458723.
3. A. Koenecke, A. Nam, E. Lake, et al. Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14), 7684–7690. (2020)
4. A. Akera et al., How much speech data is necessary for ASR in African languages? An evaluation of data scaling in Kinyarwanda and Kikuyu, in Proc. Int. Conf. Learn. Represent., Kigali, Rwanda, pp. 1-15. (2025)
5. A. Nagrani et al., The People's Speech: A Large-Scale Diverse English Speech Recognition Dataset for Commercial Use, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6162-6166, (2022). doi: 10.1109/ICASSP43922.2022.9746470.
6. R. Ardila et al., Common Voice: A Massively-Multilingual Speech Corpus, in Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, pp. 4211-4215. (2020)
7. A. Conneau et al., Unsupervised Cross-lingual Representation Learning for Speech Recognition, in Proc. Interspeech 2021, Brno, Czech Republic, pp. 2446-2450. (2021). doi: 10.21437/Interspeech.2021-329.
8. K. Sekoyan et al., Granary: Speech Recognition and Translation Dataset in 25 European Languages. arXiv:2505.13404v1. (2025)