NUMB3RS BENCHMARK: FILLING THE GAP IN SPEECH RECOGNITION OF NUMERALS

Anonymous ACL submission

Abstract

Standard Word error rate (WER) metric is measuring the proportion of substitutions, insertions, and deletions relative to the reference transcript. However, it treats all tokens equally, even though numbers carry much higher risk when misrecognized. A single incorrect digit can change an address, payment amount, or date, yet existing automatic speech recognition (ASR) benchmarks rarely provide systematic evaluation of numerical expressions. We introduce the **Numb3rs Benchmark**, a dataset designed specifically for assessing recognition of numbers in both written and spoken form. It is derived from Text Normalization corpus, which contains diverse numeric expressions (e.g., dates, times, money, telephone numbers, measurements) paired with normalized text [example pair: (10, ten)]. We extend this resource with speech generated through text-to-speech synthesis and afterwards validated by human annotation. Experiments show that state-of-the-art ASR systems achieve only about 15% WER on this benchmark—compared to $\sim$3% on LibriSpeech—demonstrating that number recognition remains a significant challenge.

1 Introduction

Automatic Speech Recognition (ASR) has seen rapid advances in recent years, with systems now achieving strong performance on widely used benchmarks.[1] These improvements have enabled ASR to become a core technology in applications ranging from voice assistants and call centers to navigation systems and healthcare.

Despite this progress, ASR errors remain highly consequential in practice. While standard evaluation relies on word error rate (WER)—which measures the proportion of substitutions, insertions, and deletions relative to the reference transcript—this metric treats all tokens equally. In many scenarios, however, some tokens carry far greater risk than others. Numerical expressions are a prime example: a single misrecognized digit can change a payment amount, misroute a delivery address, or alter a date or time. In high-stakes domains such as finance or medicine, these mistakes can have severe consequences.

Existing benchmarks do not systematically capture this problem. Some corpora contain limited digit-only material (Leonard and Doddington, 1993; Jackson et al., 2017; Becker et al., 2023), while others include numbers incidentally but without clear normalization conventions (Hernandez et al., 2018; Conneau et al., 2022). Recent work has addressed text normalization challenges (Sproat and Jaitly, 2017; Huber and Waibel, 2025; Zhang et al., 2021), but systematic evaluation of numeric recognition in ASR remains limited. As a result, ASR models are trained and evaluated on a mixture of normalized (*123*) and inverse text-normalized (*one hundred twenty-three*) forms, leading to inconsistent predictions and underreporting of numerical errors.

During the evaluation of a newly trained ASR model, we observed systematic inconsistencies in the transcription of numeric expressions. For instance, some years were produced in written (digit) form while adjacent ones appeared in spoken form, and long monetary values occasionally contained missing digits. These irregularities highlighted the lack of a dedicated benchmark for assessing numeric consistency, which ultimately motivated the creation of our dataset.

In this work, we introduce a dedicated evaluation set for number recognition in ASR - **Numb3rs Benchmark**[2]. It features the following key advancements.

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

[2] https://anonymous.4open.science/r/OpenDS-5E69

- Created using Google's Text Normalization corpus, it provides audio and text representation in both normalized and inverse-normalized conditions.

- The benchmark encompasses 12 balanced semiotic groups of numeric expressions, including dates, times, money, measurements, telephone numbers, and addresses.

- Evaluation enables us to rank state-of-the-art systems by achieving error rates ranging from 9% to 37% on this benchmark.

2 Benchmark Creation

The Numb3rs Benchmark was constructed in three main stages: extracting numeric expressions, generating spoken realizations, and verifying alignment through human annotation. Each step was designed to ensure coverage of realistic numerical categories while maintaining audio quality and annotation reliability.

2.1 Extraction of Numeric Entities

We began with the Google Text Normalization corpus (Sproat and Jaitly, 2017), a large-scale resource mapping raw text to normalized forms. All entities containing at least one numeric token were initially included. During inspection, we observed that certain categories could not be naturally vocalized. For example, website addresses are spelled out letter by letter in the source corpus, a form rarely used in spoken communication. Similarly, the *verbatim* category, which preserves all original formatting, often produced unnatural utterances. These categories were relatively small and were excluded from the benchmark. The remaining data was grouped into categories such as dates, times, money, telephone numbers, measurements, fractions, ordinals, and cardinals.

2.2 Speech Generation

Textual utterances were synthesized into audio using the Magpie text-to-speech (TTS) system (NVIDIA, 2025), with one of six built-in voices randomly chosen for each utterance. The output was generated at 44 kHz and used in this form during annotation. For experimental evaluation, all audio was downsampled to 16 kHz mono, matching the requirements of most ASR systems. No additional post-processing (e.g., noise filtering or silence trimming) was applied, as Magpie output provided sufficiently clean speech. Problematic or artifact-heavy cases were filtered out during the verification stage.

Example utterances across categories (TN vs. ITN).

DATE
TN: 2025-09-24
ITN: September twenty-fourth, twenty twenty-five

MONEY
TN: $1,250
ITN: one thousand two hundred fifty dollars

TIME
TN: 07:45
ITN: seven forty-five

TELEPHONE
TN: 555-3981
ITN: five five five, three nine eight one

FRACTION
TN: 3/4
ITN: three quarters

DECIMAL
TN: 12.07
ITN: twelve point zero seven

Figure 1: Example utterances across categories, with TN forms in green and ITN forms in blue.

2.3 Human Verification

All generated audio underwent human verification to ensure alignment with the intended transcript. Annotators were given explicit guidelines:

- If minor artifacts were present but the entity remained clearly recognizable, the utterance was retained.

- If artifacts obscured recognition, the utterance was discarded.

- If the spoken audio did not match the text, the transcript was corrected to reflect what was actually pronounced, provided that it was still semantically consistent with the original.

Ambiguous cases, such as expressions like "twenty to ten," which can be interpreted as either *22:10* or *20 to 10*, were excluded, since even human listeners struggled to agree on an interpretation. Annotation was performed in a single pass, with 20% of utterances double-checked by a second group for quality assurance. In total 3k utterances was filtered out for one of the above reasons.

3 Dataset Description

The final Numb3rs Benchmark contains 10.1k utterances (approximately 5 hours of speech) covering a wide range of numeric categories. Each category reflects frequent real-world usage, ensuring that the benchmark provides diverse and representative challenges for ASR models. As the benchmark is designed exclusively for evaluation, no train/dev/test splits are provided. Instead, the entire set is intended to serve as a test resource for measuring ASR accuracy on numerical expressions. Table 1 summarizes the distribution of categories by total hours and percentage share of the dataset.

Table 1: Category distribution of the Numb3rs Benchmark. Durations are given in hours, with percentage share of total audio.

Category	Hours	Share (%)
TELEPHONE	1.02	20.9
DATE	0.51	10.4
MEASURE	0.46	9.4
MONEY	0.45	9.1
DECIMAL	0.42	8.5
TIME	0.40	8.2
FRACTION	0.39	8.0
ADDRESS	0.31	6.4
DIGIT	0.30	6.1
CARDINAL	0.24	4.9
ORDINAL	0.24	4.9
PLAIN	0.16	3.2
Total	**4.89**	**100.0**

3.1 Qualitative Examples

To illustrate the diversity of the benchmark, Figure 1 presents example utterances from several categories. Both text-normalized (TN) and inverse text-normalized (ITN) forms are included, reflecting the two evaluation conditions.

By including both TN and ITN forms, the benchmark enables measurement of ASR accuracy under two common conventions. This dual representation allows researchers to study not only raw recognition accuracy but also normalization consistency, an often-overlooked source of errors in production systems.

4 Experiments

We evaluated the Numb3rs Benchmark using top-performing ASR models from the Hugging Face ASR leaderboard. We tested FastConformer (Rekesh et al., 2023), multiple Parakeet variants and Canary (Puvvada et al., 2024), Voxtral (Liu et al., 2025), Phi-4 (Abdin et al., 2024), Whisper (Radford et al., 2022), and SeamlessM4T (Communication et al., 2023). To ensure comparability, the evaluation strictly followed the normalization procedure implemented in the leaderboard repository.[3] As a result, the reported word error rates (WER) are directly comparable to leaderboard scores and would be identical if this dataset were officially integrated into the leaderboard.

Table 2 presents the evaluation results of state-of-the-art systems on the newly proposed benchmark. Despite strong leaderboard performance on general ASR tasks, even the best models produce substantial error rates (10–20%), while others degrade sharply, exceeding 30%.

Table 2: WER (%) on the Numb3rs Benchmark.

Model	WER (%)
stt_en_fastconformer_hybrid_large	14.43
parakeet-tdt_ctc-110m	17.33
parakeet-rnnt-1.1b	11.34
canary-1b	11.62
parakeet-tdt-0.6b-v2	16.91
Voxtral-Mini-3B-2507	31.87
Phi-4-multimodal-instruct	9.85
whisper-large-v3-turbo	37.3
seamless-m4t-v2-large	26.64

4.1 Analysis

The results show that even leaderboard-leading models such as Parakeet-RNNT, Canary, and Phi-4 exhibit WER between 9.85–11.62% on numerical expressions. This contrasts sharply with their performance on general corpora (3–6% WER). Smaller or less specialized systems (Parakeet-TDT-110m, Voxtral-Mini, Whisper-large-v3-turbo) fare much worse, with error rates exceeding 17–37%. These findings underscore the persistent difficulty of number recognition, even in state-of-the-art ASR models.

To understand which numeric categories present the greatest challenge, we analyzed the best-performing model (Phi-4) across all semantic categories. Table 3 shows substantial variation in performance. Cardinal numbers achieve near-

[3] https://github.com/huggingface/open_asr_leaderboard

perfect recognition (1.67% WER), while fractions and plain text categories prove far more difficult (26.43% and 27.86% respectively).

Table 3: Category-level WER (%) for Phi-4, the best-performing model on Numb3rs Benchmark. Performance varies dramatically across numeric categories.

Category	WER (%)
CARDINAL	1.67
MONEY	2.05
MEASURE	3.48
DECIMAL	4.51
DATE	5.95
DIGIT	6.23
TELEPHONE	7.77
ORDINAL	9.70
ADDRESS	9.79
TIME	9.81
FRACTION	26.43
PLAIN	27.86

4.2 Controlled Fine-Tuning Experiments

In addition to zero-shot evaluation, we conducted fine-tuning experiments with canary-1b. The model was fine-tuned on the same training corpora as its original release, with variations in how numerical expressions were represented. Evaluation was performed on the Numb3rs Benchmark without HF normalization to separately assess TN and ITN performance. Table 4 summarizes the results. In this setup, the benchmark was divided into two equal parts, and evaluation was performed on the second half only.

Table 4: WER (%) on Numb3rs Benchmark after fine-tuning `canary-1b`. Format: TN / ITN.

Training Condition	TN	ITN
Benchmark split (50% train)	1.27	1.54
Corpora undergo TN	8.7	19.4
(TN + ITN)	1.7	3.2

Training on first half of the benchmark yields near-perfect performance (1.27% TN, 1.54% ITN), confirming the dataset is learnable with direct supervision. Normalization-specific training (Experiment 2) improved performance on the corresponding text form but degraded on the opposite form. In these experiments, corpora were normalized using NVIDIA NeMo Text Normalization., while bilingual training (experiment 4) with both TN and ITN representations provides more balanced results. These findings suggest that explicit exposure to both normalization forms during training is necessary for robust numeric ASR across different output conventions.

4.3 Findings

These experiments highlight that fine-tuned models can adapt well to the benchmark when explicitly exposed to numbers, but generalization across normalization forms is limited. Treating TN and ITN as parallel conditions during training reduces this gap, suggesting that bilingual-style training is a promising strategy for robust numerical ASR.

5 Conclusion

In this work, we present the **Numb3rs Benchmark**, a dedicated evaluation set for number recognition in ASR. Derived from Google's Text Normalization corpus and extended with speech generated by text-to-speech synthesis and verified by human annotation, the benchmark covers a diverse range of numeric expressions such as dates, times, money, measurements, telephone numbers, and addresses. It enables evaluation in both normalized and inverse-normalized conditions, providing a systematic testbed for studying ASR performance on numbers.

Our experiments show that state-of-the-art systems, despite their strong leaderboard performance, still reach error rates of around 15% on this benchmark, demonstrating that number recognition remains a significant open challenge.

6 Limitations

This benchmark is based on synthesized speech from a single TTS system (Magpie), which may not capture the full variability of human speech patterns. The numeric expressions originate from written text, potentially omitting naturally occurring spoken formulations. Additionally, the benchmark focuses solely on English and does not yet address multilingual numeric recognition. Finally, the benchmark integrates data sources that are either distributed under the *Creative Commons Attribution 4.0 (CC BY 4.0)* license or explicitly approved for research use; as such, redistribution must comply with these original licensing terms.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. 2023. Audiomnist: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *Preprint*, arXiv:1807.03418.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, and 49 others. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *Preprint*, arXiv:2308.11596.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *Preprint*, arXiv:2205.12446.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. *TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation*, page 198–208. Springer International Publishing.

Christian Huber and Alexander Waibel. 2025. Handling numeric expressions in automatic speech recognition. *Preprint*, arXiv:2408.00004.

Zohar Jackson, Hereman Nicolas, and César Souza. 2017. Jakobovski/free-spoken-digit-dataset v1.0.5. Free Spoken Digit Dataset (FSDD), version 1.0.5.

R. Gary Leonard and George Doddington. 1993. TIDIGITS ldc93s10. Web Download. Catalog No. LDC93S10.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, and 87 others. 2025. Voxtral. *Preprint*, arXiv:2507.13264.

NVIDIA. 2025. Enhancing multilingual human-like speech and voice cloning with NVIDIA Riva TTS. NVIDIA Technical Blog.

Krishna C. Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, Jagadeesh Balam, and Boris Ginsburg. 2024. Less is more: Accurate speech recognition & translation without web-scale data. *Preprint*, arXiv:2406.19674.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. *Preprint*, arXiv:2305.05084.

Richard Sproat and Navdeep Jaitly. 2017. Rnn approaches to text normalization: A challenge. *Preprint*, arXiv:1611.00068.

Yang Zhang, Evelina Bakhturina, Kyle Gorman, and Boris Ginsburg. 2021. Nemo inverse text normalization: From development to production. *Preprint*, arXiv:2104.05055.