Soniox Italian Speech Recognition Benchmarks

Soniox Inc, October 2023
support@soniox.com
https://soniox.com

Abstract

Soniox has conducted an extensive evaluation of the accuracy of different speech recognition providers in the industry. The benchmarks are summarized as follows:

- **Providers evaluated:** Soniox, OpenAI, Google, AWS, Azure, NVIDIA Riva, Deepgram, AssemblyAI, Speechmatics, and Rev AI.
- **Languages evaluated:** Italian.
- **Evaluation datasets**: real-world dataset of YouTube videos varying in acoustic conditions, speaking styles, accents, topics, and the number of speakers.
- **Ground truth transcriptions** were transcribed and double-reviewed by humans, then normalized to ensure a fair evaluation across different providers.
- **Processing modes evaluated**: asynchronous transcription (file or batch) and streaming (live) transcription.
- **Results**:
    - Soniox achieved the most accurate speech recognition results in both async and streaming modes by large margins.
    - In async mode, AssemblyAI holds second place, but Soniox surpasses it with **41% greater accuracy**. In streaming mode, both Azure and Google tie for second, yet Soniox leads with over **48% higher accuracy**.
    - The lowest accuracy was observed with AWS, NVIDIA, and Rev AI.

Results

Word Error Rates (WER)

For each dataset, we evaluated all providers in **asynchronous** (file or batch) and **streaming** (live) modes. The metric reported is Word Error Rate (WER), an industry standard metric for evaluating the accuracy of speech recognition systems, which measures the percentage of words misrecognized. A lower WER indicates a more accurate speech recognition system.

Dataset	Hours	Mode	Soniox	OpenAI	Google	AWS	Azure	NVIDIA	Deepgram	AssemblyAI	Speechmatics	Rev AI
YouTube videos	10	Async	**4.81%**	8.41%	16.07%	8.87%	8.54%	12.24%	10.13%	6.74%	7.09%	9.07%
		Streaming	**5.43%**	N/A	8.76%	18.13%	8.54%	13.95%	12.86%	N/A	9.49%	18.13%

Soniox achieved the lowest WER (highest accuracy) in both modes by extremely large margins. In async mode, the second place belongs to AssemblyAI with 1.93% higher WER. In streaming mode, Azure/Google placed second with more than 3.10% higher WER. All other providers fall far behind Soniox and incur much higher WERs.

Note that OpenAI (Whisper) and AssemblyAI are only supported in Async mode and not in Streaming mode.

Soniox Improvement

The table below shows how much of an improvement Soniox speech recognition AI brings over other providers, if you were to transcribe the same audio.

Dataset	Hours	Mode	Soniox	OpenAI	Google	AWS	Azure	NVIDIA	Deepgram	AssemblyAI	Speechmatics	Rev AI
YouTube videos	10	Async	N/A	56%	80%	59%	57%	73%	65%	41%	45%	60%
		Streaming	N/A	N/A	49%	79%	48%	71%	68%	N/A	54%	79%

Soniox outperforms other providers by extremely large margins. For example, in streaming mode, **Soniox makes 48% fewer errors than Azure/Google**. That is, Soniox fixes 48% of the errors that Azure's/Google's speech recognition AI makes.

Note, the improvement metric has been computed by taking the original WER reported in the table above, from which we subtracted the estimated amount of error in ground truth transcriptions. We were conservative in these estimates and estimated only 2% of human error for the YouTube videos dataset.

Methodology

Datasets

To represent **real-world** speech recognition use cases, we selected our evaluation dataset to include difficult real-world speech environments, conversations with crosstalk and interjections. All datasets are spoken in the Italian language.

Dataset	Difficulty Level	Example	Duration
YouTube videos	Medium / High	Live Cooking Product Review Conversations	10 hours

Metric

We calculated the Word Error Rate (WER) following the standard definition:

$$WER = \frac{Number\ of\ Words\ Recognized\ Incorrectly}{Number\ of\ Words\ in\ Ground\ Truth}$$

$$Number\ of\ Words\ Recognized\ Incorrectly = Substitutions + Insertions + Deletions$$

Ground Truth Transcriptions

1. Ground truth transcriptions were labeled and double-reviewed by humans.
2. Ground truth and provider transcriptions were then normalized to achieve a fair evaluation across different providers. See examples below.

Normalization Type	Normalization Performed
Filler words	Filler words (e.g. uhm) were removed
Punctuations	Punctuations were removed
Capitalization	Catarina ⇒ catarina

Models Evaluated

We evaluated the latest and most accurate models from each provider in both streaming and asynchronous processing modes. To do so, we integrated with every provider's API carefully following their documentation. All models were evaluated on October 24th, 2023.

Provider	**Async Model**	**Streaming Model**
Soniox	it_v2	it_v2_lowlatency
OpenAI Whisper	large-v2 (it)	Streaming is not supported
Google	latest_long (it-IT)	latest_long (it-IT)
AWS	Amazon Transcribe (it-IT)	Amazon Transcribe (it-IT)
Azure	Standard (it-IT)	Standard (it-IT)
NVIDIA Riva	Riva ASR (it-IT)	Riva ASR (it-IT)
Deepgram	General Enhanced (it)	General Enhanced (it)
AssemblyAI	Asynchronous Transcription (it)	Streaming is not supported
Speechmatics	Transcribe a File Enhanced (it)	Transcribe in Real-time Enhanced (it)
Rev AI	Asynchronous Speech-to-Text API (it)	Streaming Speech-to-Text API (it)

Contact Information

Email: support@soniox.com

If you have any questions, comments or suggestions about the benchmarks, feel free to reach out to us. We welcome your feedback and are always looking for ways to improve the benchmarks.

If you would like us to include a new evaluation dataset in the next benchmarking report, also reach out to us and we would be happy to do so.

Benchmark Your Dataset

If you would like to evaluate speech recognition providers on your dataset, please reach out to us, and we will assess the providers of interest on your dataset and provide you with the complete results, including the outputs of every provider.