Soniox English Speech Recognition Benchmarks

Soniox Inc, September 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:** English.
- **Evaluation datasets**: real-world datasets 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 on all datasets in both async and streaming modes.
    - Second place was shared by OpenAI, Speechmatics, and AssemblyAI; however, there was a significant accuracy gap between Soniox and the second-place winners. For example, **Soniox achieved a 24% higher accuracy** rate than OpenAI on YouTube videos.
    - The lowest accuracy was observed with Google, Deepgram, and AWS.

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	**4.52%**	5.32%	10.55%	7.47%	7.58%	9.42%	8.96%	6.43%	5.88%	7.59%
		Streaming	**5.18%**	N/A	9.45%	10.58%	7.59%	10.11%	11.75%	7.91%	6.27%	9.07%
Conversations	10	Async	**5.64%**	9.31%	15.12%	11.77%	12.41%	11.51%	13.79%	8.99%	10.10%	12.25%
		Streaming	**6.82%**	N/A	14.32%	15.96%	12.41%	12.62%	17.34%	11.44%	9.71%	13.39%

Soniox achieved the lowest WER (highest accuracy) in both modes and on both datasets, followed by OpenAI, Speechmatics and Assembly AI. The difference in accuracy between Soniox and any second place provider is large. All other providers fall far behind Soniox and incur much higher WERs on these real-world audio datasets.

Note that OpenAI (Whisper) is 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	24%	71%	54%	55%	66%	64%	43%	35%	55%
		Streaming	N/A	N/A	57%	63%	43%	61%	67%	46%	26%	55%
Conversations	10	Async	N/A	58%	78%	70%	72%	69%	76%	56%	63%	71%
		Streaming	N/A	N/A	66%	71%	59%	60%	73%	55%	43%	63%

Soniox outperforms other providers by large margins. For example, in Async mode on the *Conversations* dataset, **Soniox makes 58% fewer errors than OpenAI**. That is, Soniox “fixes” 58% of the errors that OpenAI's speech recognition AI makes. Similar numbers apply to other providers.

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 and 3% of error for the *Conversations* dataset.

Methodology

Datasets

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

Dataset	Difficulty Level	Example	Duration
YouTube videos	Medium	MIT Open Course	10 hours
Conversations	Medium / High	Panel Discussion Various 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 August 29th, 2023.

Provider	**Async Model**	**Streaming Model**
Soniox	en_v2	en_v2_lowlatency
OpenAI Whisper	medium.en	Streaming is not supported
Google	Enhanced Video	Enhanced Video
AWS	Amazon Transcribe	Amazon Transcribe
Azure	Standard	Standard
NVIDIA Riva	Riva ASR	Riva ASR
Deepgram	Nova General	Nova General
AssemblyAI	Asynchronous Transcription	Real-Time Streaming Transcription
Speechmatics	Transcribe a File Enhanced	Transcribe in Real-time Enhanced
Rev AI	Asynchronous Speech-to-Text API	Streaming Speech-to-Text API

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.