Soniox Speech-To-Text Benchmarks

Soniox Inc, November 2022
v2, updated 2022-12-05
support@soniox.com
https://soniox.com

Abstract

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

- **Providers evaluated**: Soniox, Google, AWS, Azure, Rev AI, Deepgram, AssemblyAI, OpenAI Whisper and Speechmatics.
- **Processing modes evaluated**: asynchronous transcription (file) and streaming transcription.
- **Evaluation datasets**: 4 different real-world datasets varying in acoustic conditions, speaking styles, accents and topics in the English language.
- **Ground truth transcriptions** were transcribed and double-reviewed by humans then normalized to ensure a fair evaluation across different providers.
- **Results**:
    - Overall, Soniox achieved the most accurate speech recognition results in both async and streaming modes across all 4 datasets, followed by Azure and Speechmatics. The lowest overall performance was obtained by Deepgram, AssemblyAI and Google. In the middle of the pack are Rev AI and AWS.
    - In streaming mode, Soniox leads with a wider margin compared to other providers. AssemblyAI and Deepgram had the lowest performance in the streaming mode.
- The benchmarks were conducted with a high level of professionalism. Hundreds of hours of human time were invested to develop a benchmarking framework that tries to fairly evaluate the accuracy of different speech-to-text providers.

Results

Word Error Rates (WER)

The following table contains the benchmark results. For each dataset, we evaluated all providers in asynchronous (file) and streaming processing 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-to-text system.

Dataset	Hours	Mode	Soniox	Google	AWS	Azure	Rev AI	Deepgram	AssemblyAI	OpenAI	Speechmatics	Date
News Reporting and Broadcasting	10	Async	2.05%	5.00%	3.63%	2.99%	5.71%	5.12%	3.58%	10.50%	2.27%	2022-11-22*
		Streaming	2.05%	4.99%	5.22%	2.98%	8.92%	7.08%	8.68%	N/A	3.00%	2022-11-22*
Video Lectures & Education	10	Async	4.43%	7.73%	5.94%	5.80%	6.15%	6.72%	6.01%	6.44%	5.40%	2022-11-21*
		Streaming	4.43%	7.71%	8.83%	5.81%	8.71%	9.27%	9.85%	N/A	5.73%	2022-11-21*
Conversations with Crosstalk and Interjections	10	Async	6.76%	11.64%	6.85%	7.72%	6.69%	11.25%	9.83%	7.30%	7.61%	2022-11-23*
		Streaming	6.76%	11.64%	10.70%	7.72%	9.48%	14.42%	15.09%	N/A	7.81%	2022-11-23*
Background Noise, Crosstalks, Unclear Speech	10	Async	14.83%	23.00%	18.07%	16.38%	15.82%	21.68%	19.00%	15.94%	16.95%	2022-11-23*
		Streaming	14.83%	22.95%	22.84%	16.37%	19.74%	25.27%	25.02%	N/A	17.42%	2022-11-23*

*Speechmatics benchmarks were run on 2022-12-02/04.

Observations

Overall Performance

	Soniox	Google	AWS	Azure	Rev AI	Deepgram	AssemblyAI	OpenAI	Speechmatics
Average WER (async + streaming)	7.02%	11.83%	10.26%	8.22%	10.15%	12.60%	12.13%	N/A	8.27%

Overall, Soniox achieved the lowest average WER of 7.02%, followed by Azure and Speechmatics with an average WER of 8.22% and 8.27%, respectively. Deepgram, AssemblyAI and Google generated the highest average WERs ~12%, trailing Soniox by ~5% absolute. In the middle of the pack are Rev AI and AWS with an average WERs ~10%, trailing Soniox by ~3% absolute. Note that OpenAI could not be considered on an overall level due to their unsupported API for streaming mode.

Asynchronous (File) Performance

	Soniox	Google	AWS	Azure	Rev AI	Deepgram	AssemblyAI	OpenAI	Speechmatics
Average WER (async)	7.02%	11.84%	8.62%	8.22%	8.59%	11.19%	9.61%	10.05%	8.06%

In the asynchronous mode, Soniox generated the lowest average WER of 7.02%, followed by Speechmatics, Azure, Rev AI and AWS with average WERs ~ 8%. Google and Deepgram yielded the highest average WERs of 11.84% and 11.19%, respectively. In the middle, AssemblyAI and OpenAI have WERs of ~10%.

Streaming (Real-Time) Performance

	Soniox	Google	AWS	Azure	Rev AI	DeepGram	AssemblyAI	OpenAI	Speechmatics
Average WER (streaming)	7.02%	11.82%	11.90%	8.22%	11.71%	14.01%	14.66%	N/A	8.49%

In the streaming mode, Soniox generated the lowest average WER of 7.02%. Following Soniox are Azure and Speechmatics with an average WER of ~8%. Deepgram and AssemblyAI yielded the highest average WER of ~14%. In the middle of the pack are Rev AI, Google and AWS with average WERs of ~12%.

Streaming vs Async Performance Gaps

	Soniox	Google	AWS	Azure	Rev AI	Deepgram	AssemblyAI	OpenAI	Speechmatics
Diff (Streaming - Async) Average WER	0.00%	-0.02%	3.28%	0.00%	3.12%	2.82%	5.06%	N/A	0.43%

When comparing the WER difference between streaming vs async performances (e.g. Azure streaming WER vs Azure async WER) , we noticed that for Soniox, Google and Azure there is virtually no difference in WERs between the streaming and async modes. However, this is not the case for AssemblyAI, AWS, Rev AI and Deepgram, where the difference in WER between streaming and async is ~ 3 to 5% absolute. Speechmatics had a moderate streaming-to-async accuracy difference of 0.43%.

OpenAI Insertion Errors

We saw that OpenAI had a high insertion error rate and frequently inserted extra words at the end of audio segments. Please see an example below.

Example: OpenAI often inserts extra words at the end of a sentence resulting in high insertion error rate.

	Transcription
Ground Truth	he also offered to pardon islamist fighters who surrender the
OpenAI	he also offered to pardon islamist fighters who surrender he **also has been rem recording**

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 in addition to cleaner audio, such as news reporting and educational lectures. All datasets are spoken in the English language.

Dataset	Difficulty Level	Example	Duration
News Reporting and Broadcasting	Low	NBC Nightly News	10 hours
Video Lectures & Education	Low-Medium	MIT Open Course	10 hours
Conversations with Crosstalk and Interjections	Medium	Panel Discussion	10 hours
Background Noise, Crosstalks or Unclear Speech	High	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
Numbers	21 ⇒ twenty one
Metrics	mg ⇒ milligrams
Brand names	23AndMe ⇒ twenty three and me

Contractions	gonna ⇒ going to
Filler words	Filler words (e.g. um) were removed
Punctuations	Punctuations were removed
Capitalization	This ⇒ this

Models Evaluated

We evaluated the most current and accurate model(s) from each provider in both streaming and asynchronous processing modes. To do so, we integrated with every provider's API carefully following their documentation.

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

* Soniox currently uses the same model for streaming and async processing.
** We chose the *small.en* model (244M params) due to the practical feasibility (it is computationally expensive (infeasible) to run large models with 1000M+ params on large amounts of audio data).

Contact Information

Email: support@soniox.com

If you have any questions, comments or suggestions about the benchmarks, please reach out to us via email. 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, please reach out to us and we would be happy to do so.

Benchmark Your Dataset

If you would like to evaluate speech-to-text providers on your dataset, feel free to reach out to us and we will evaluate providers of interest on your dataset and return back to you the full results, including the outputs of every provider.