Soniox Speech Recognition Benchmarks

German and Spanish Languages

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

Abstract

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

- **Providers evaluated**: Soniox, Google, AWS, Azure, Rev AI, Deepgram, AssemblyAI, OpenAI Whisper, Speechmatics and NVIDIA Riva.
- **Languages evaluated:** German and Spanish
- **Evaluation datasets**: real-world datasets varying in acoustic conditions, speaking styles, accents and topics.
- **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) and streaming transcription.
- **Results**:
    - Overall, Soniox achieved the most accurate speech recognition results in both async and streaming modes across all German and Spanish datasets. The second place belongs to Speechmatics (German) and Azure (Spanish).
    - Soniox achieved 23% higher accuracy on Spanish and 27% higher accuracy on German compared to the second place provider, i.e. about every 4th word misrecognized by Speechmatics or Azure, was correctly recognized by Soniox.
    - The lowest overall performance was obtained by Google and AWS. In the middle of the pack are the remaining providers.
- The benchmarks were conducted with a high level of professionalism. We invested significant engineering resources to develop a benchmarking framework that tries to fairly evaluate the accuracy of different speech recognition 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 recognition system.

Language	Dataset	Hours	Mode	Soniox	Google	AWS	Azure	Rev AI	Deepgram	AssemblyAI	OpenAI	Speechmatics	NVIDIA
German	**Video Lectures & Education**	10	Async	5.92%	16.50%	15.16%	9.40%	16.85%	10.03%	13.61%	14.45%	8.13%	10.45%
			Streaming	5.92%	15.89%	16.22%	9.41%	16.22%	12.84%	N/A	N/A	8.98%	12.16%
Spanish	**Video Lectures & Education**	10	Async	4.49%	19.60%	12.08%	6.40%	11.01%	12.29%	9.93%	14.61%	7.38%	9.53%
			Streaming	4.49%	17.48%	12.39%	6.40%	13.51%	13.14%	N/A	N/A	7.87%	10.19%
	Conversations with Crosstalk and Interjections	10	Async	10.84%	36.21%	25.94%	13.59%	17.17%	17.55%	19.77%	18.83%	13.91%	17.85%
			Streaming	10.84%	34.39%	27.27%	13.58%	20.07%	21.12%	N/A	N/A	14.39%	18.62%

Analysis

German Language

Language	Dataset	Mode	Soniox	Google	AWS	Azure	Rev AI	Deepgram	AssemblyAI	OpenAI	Speechmatics	NVIDIA
German	**Video Lectures & Education**	Async	5.92%	16.50%	15.16%	9.40%	16.85%	10.03%	13.61%	14.45%	8.13%	10.45%
		Streaming	5.92%	15.89%	16.22%	9.41%	16.22%	12.84%	N/A	N/A	8.98%	12.16%
		Async+Streaming	5.92%	16.20%	15.69%	9.41%	16.54%	11.44%	N/A	N/A	8.56%	11.31%

Overall (Async+Streaming), Soniox achieved the lowest average WER of 5.92%, followed by Speechmatics with WER of 8.56% and Azure with WER of 9.41%. The highest WERs were achieved by Rev AI, Google and AWS with WERs ~16%. NVIDIA and Deepgram are in the middle with the WERs ~11%.

AssemblyAI and OpenAI do not support streaming API for German language and could not be considered in the overall evaluation.

In the asynchronous mode, Soniox generated the lowest average WER of 5.92%, followed by Speechmatics (8.13%), Azure (9.40%), Deepgram (10.03%), NVIDIA (10.45%), AssemblyAI (13.61%), OpenAI (14.45%), AWS (15.16%), Google (16.50%) and Rev AI (16.85%).

In the streaming mode, Soniox also generated the lowest average WER of 5.92%, followed by Speechmatics (8.98%), Azure (9.41%), NVIDIA (12.16%), Deepgram (12.84%), AWS and Rev AI (16.22%), and Google (15.89%).

Spanish Language

Language	Dataset	Mode	Soniox	Google	AWS	Azure	Rev AI	Deepgram	AssemblyAI	OpenAI	Speechmatics	NVIDIA
Spanish	**Video Lectures & Education**	Async	4.49%	19.60%	12.08%	6.40%	11.01%	12.29%	9.93%	14.61%	7.38%	9.53%
		Streaming	4.49%	17.48%	12.39%	6.40%	13.51%	13.14%	N/A	N/A	7.87%	10.19%
		Async+Streaming	4.49%	18.54%	12.24%	6.40%	12.26%	12.72%	N/A	N/A	7.63%	9.86%
	Conversations with Crosstalk and Interjections	Async	10.84%	36.21%	25.94%	13.59%	17.17%	17.55%	19.77%	18.83%	13.91%	17.85%
		Streaming	10.84%	34.39%	27.27%	13.58%	20.07%	21.12%	N/A	N/A	14.39%	18.62%
		Async+Streaming	10.84%	35.30%	26.61%	13.59%	18.62%	19.34%	N/A	N/A	14.15%	18.24%
	Merged Datasets	Async	7.67%	27.91%	19.01%	10.00%	14.09%	14.92%	14.85%	16.72%	10.65%	13.69%
		Streaming	7.67%	25.94%	19.83%	9.99%	16.79%	17.13%	N/A	N/A	11.13%	14.41%
		Async+Streaming	7.67%	26.92%	19.42%	9.99%	15.44%	16.03%	N/A	N/A	10.89%	14.05%

Overall, Soniox achieved the lowest average WER of 7.67%, followed by Azure with WER of 9.99% and Speechmatics with WER of 10.89. Google's performance on both Spanish datasets was disappointing, with an average WER of 26.92%. In the middle we have NVIDIA, Rev AI and Deepgram with an average WER in the range between 14% to 16%.

AssemblyAI and OpenAI do not support streaming API for Spanish language and could not be considered in the overall evaluation.

In the asynchronous mode, Soniox generated the lowest average WER of 7.67%, followed by Azure (10.00%), Speechmatics (10.65%), NVIDIA (13.69%), Rev AI (14.09%), AssemblyAI (14.85%), Deepgram (14.92%), OpenAI (16.72%), AWS (19.01%), and Google (27.91%).

In the streaming mode, Soniox also generated the lowest average WER of 7.67%, followed by Azure (9.99%), Speechmatics (11.13%), NVIDIA (14.41%), Rev AI (16.79%), Deepgram (17.13%), AWS (19.83%), and Google (25.94%).

Streaming vs Async Performance

	Language	Soniox	Google	AWS	Azure	Rev AI	Deepgram	AssemblyAI	OpenAI	Speechmatics	NVIDIA
Differance (Streaming - Async)	German	0.00%	-0.61%	1.06%	0.01%	-0.63%	2.81%	N/A	N/A	0.85%	1.71%
	Spanish	0.00%	-1.97%	0.82%	0.00%	2.70%	2.21%	N/A	N/A	0.49%	0.72%

When comparing the WER difference between streaming vs async performances (e.g. Azure streaming WER vs Azure async WER) , we noticed that for Soniox and Azure there is virtually no difference in WERs between their streaming and async modes. However, this is not the case for other providers. For example, the Deepgram difference in WER between streaming and async is 2.81% for German and 2.21% for Spanish. NVIDIA also has a reasonably high difference in streaming-to-async WER of 1.71% on the German dataset.

AssemblyAI and OpenAI could not be considered on streaming vs async evaluation, since they do not support streaming API for these languages.

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, and educational lectures. All datasets are spoken in the German and Spanish language.

Dataset	Difficulty Level	Example	Duration
Video Lectures & Education	Low-Medium	MIT Open Course	10 hours
Conversations with Crosstalk and Interjections	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 Performed	
Normalization Type	**German**	**Spanish**
Numbers	21 ⇒ ein und zwanzig	45 ⇒ cuarenta y cinco
Metrics	mg ⇒ milligramm	km ⇒ kilómetro
Brand names	N26 ⇒ n sechs und zwanzig	N26 ⇒ n veinticinco
Contractions	gradeaus ⇒ geradeaus	m'hija ⇒ mi hija

Filler words	Filler words (e.g. ähm) were removed	Filler words (e.g. um) were removed
Punctuations	Punctuations were removed	Punctuations were removed
Capitalization	Haus ⇒ haus	España ⇒ españa

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. All models were evaluated from March 15th to March 17th.

Provider	**Async Model**	**Streaming Model**
Soniox	Precision* (de_precision, es_precision)	Precision* (de_precision, es_precision)
Google	Default (de-DE, es-ES)	Default (de-DE, es-ES)
AWS	Amazon Transcribe (de-DE, es-ES)	Amazon Transcribe (de-DE, es-US)
Azure	Standard (de-DE, es-ES)	Standard (de-DE, es-ES)
Rev AI	Asynchronous Speech-to-Text API (de, es)	Streaming Speech-to-Text API (de, es)
Deepgram	Enhanced General (de, es)	Enhanced General (de, es)
AssemblyAI	Asynchronous Transcription (de, es)	Streaming is not supported
OpenAI Whisper	large-v2	Streaming is not supported
Speechmatics	Transcribe a File Enhanced (de, es)	Transcribe in Real-time Enhanced (de, es)
NVIDIA Riva	Riva ASR (de-DE, es-ES)	Riva ASR (de-DE, es-ES)

* Soniox currently uses the same model for streaming and async processing.

Contact Information

Email: support@soniox.com

If you have any questions, comments or suggestions about the benchmarks, please 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, please 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, 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.