Soniox and Whisper
Speech Recognition Benchmarks

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

Abstract

We conducted an extensive evaluation on word recognition accuracy of Soniox and OpenAI Whisper speech recognition AI. The benchmarks are summarized as follows:

- **Evaluation datasets**: 5 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**:
    - Soniox achieved the most accurate speech recognition results across all 5 datasets.
    - Soniox was 32.61% more accurate than Whisper, meaning that on average almost every 3rd word incorrectly recognized by Whisper was correctly recognized by Soniox.
    - Whisper sometimes had a high insertion rate and recognized (hallucinated) words not spoken in the audio. Similarly, Whisper sometimes also had a high deletion rate and did not recognize words clearly spoken in the audio.
- 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. We evaluated Soniox and Whisper on 5 different datasets. 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	Soniox WER	Whisper WER	Soniox Improvement Over Whisper	Date
News Reporting and Broadcasting	10	2.00%	10.76%	81.41%	2023-03-03
Video Lectures & Education	10	4.06%	6.94%	41.50%	2023-03-03
Conversations with Crosstalk and Interjections	10	6.30%	7.07%	10.89%	2023-03-03
Telephony, Customer-Agent Phone Calls	10	7.41%	10.35%	28.41%	2023-03-03
Background Noise, Crosstalk, Unclear Speech	10	14.35%	15.51%	7.48%	2023-03-03

Observations

Performance

Overall, Soniox achieved the average WER of 6.82%, while Whisper achieved the average WER of 10.13%. That is a 32.61% improvement in accuracy for Soniox over Whisper.

The smallest gap was achieved on the dataset with background noise, crosstalk and unclear speech, Soniox leading by 7.48% in accuracy. The largest gap was achieved on news reporting and broadcasting, Soniox leading by 81.41% in accuracy. See the cause of Whisper errors in the following sections.

Whisper Insertion Errors

We saw that Whisper had a high insertion error rate and recognized extra words that were not spoken in the audio. This happened more often with the “News Reporting and Broadcasting” and “Telephony, Customer-Agent Phone Calls” dataset. Please see a couple of examples below.

Examples: Whisper sometimes recognized or hallucinated words that were not spoken in the audio, resulting in inaccurate transcription.

	Transcription
Ground Truth	he also offered to pardon islamist fighters who surrender the
Whisper	he also offered to pardon islamist fighters who surrender he **also has been rem recording**
Ground Truth	the the buyouts compared to the
Whisper	the buyouts **are moon interest and that january tenth** compared to the **windy weather**

Whisper Deletion Errors

We also saw that Whisper had sometimes a high deletion error rate and did not recognize the words clearly spoken in the audio. Categories of such deletion errors were phone numbers, zip codes and web addresses.

Examples: Whisper sometimes did not recognize clearly spoken words in the audio, resulting in inaccurate transcription.

	Transcription
Ground Truth	**eight nine seven zero two three** one **i** am speaking with cynthia sullivan concerning
Whisper	one am speaking with cynthia sullivan concerning
Ground Truth	we're america's g-m parts center visit us on the web at **chuck hutton dot com**
Whisper	we're america's g-m parts center visit us on the web at

Methodology

Datasets

To represent real-world speech recognition use cases, we selected evaluation datasets to include difficult real-world speech environments 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
Telephony, Customer-Agent Phone Calls	Medium-High	Phone Conversations	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 used the latest models from each provider. To do so, we integrated with each provider's API following their documentation.

Provider	**Model**	**Date**
Soniox	Precision*	2023-03-03
Whisper	medium.en**	2023-03-03

* Soniox currently uses the same model for streaming and async processing. Thus you get the same quality of recognition for files as for streams.
** We chose the *medium.en* model (769M params), which is meant to be the highest accuracy Whisper model for English.

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 any speech recognition provider 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 with word-by-word analysis.