By: Momen Tarek

STT&TTS Workflow

Speech to text (STT) is an essential component for **creating voice-powered experiences** that delight users. A subset of automatic speech recognition (ASR), STT algorithms enable you to apply text-based natural language processing (NLP) techniques to a user's intentions. This makes speech-to-text perfect for use cases like:

- generating video captions
- transcribing meetings
- converting voice to plain text for analysis

However, the speech-to-text landscape has become increasingly competitive in a short amount of time. Just like with **evaluating large language models (LLMs)**, AI practitioners are under pressure to test and benchmark more and more speech-to-text models to find which ones best fit their applications.

How to Pick the Best Speech-to-Text Model

When choosing a speech-to-text model, it's important to consider factors such as:

- **Word error rate (WER)**: how accurate a model is at transcription, based on quantifying how many mistakes it makes when transcribing an audio clip.
- **Words per minute (WPM)**: how fast a model processes text, a high-impact metric if you plan to process multiple or long audio clips.
- **Cost**: how much the model costs, if applicable (paid services typically price their offering per minute of audio).
- **Multilingual support**: how many languages the model or service supports.
- **Streaming**: how well a model performs for use cases that demand near real-time transcription, such as enabling sentiment analysis in customer service environments like contact centers.

The importance of these factors will differ depending on your scenario. For example, real-time voice applications may favor higher WPM at the cost of WER or even literal cost. Likewise, multilingual support may be the most important factor when developing a voice solution for broad or international audiences.

Moreover, other factors beyond those mentioned above could become priority based on your unique needs. Diarization, the ability to identify different voices and then segment the transcript by speaker, might be mission-critical in environments like board rooms and law offices.

Methodology: How We Chose and Tested Our 10 Speech-to-Text Models

We started by selecting high-performing and popular models listed in the **Artificial Analysis Speech to Text AI Model & Provider Leaderboard**. We then chose several models available from large cloud vendors, as these may already be available to you via your cloud provider. OpenAI's Whisper was the only open-source model we tested, though half of our final test list represented Whisper-based models:

1. assemblyai-universal-2
2. azure-ai-speech
3. deepgram-nova-2
4. gladia
5. groq-distil-whisper
6. groq-whisper-large-v3
7. groq-whisper-large-v3-turbo
8. openai-whisper-large-v2
9. speechmatics
10. whisper-large-v3-local

We tested the models against voice samples broken down by duration, language, and number of speakers present in the clips. This formed rough buckets of scenarios we feel users are likely to encounter in the wild.

Short-form clips, roughly 30 seconds of audio (single speaker):

- Native English speakers
- Non-native or accented English speakers
- Non-English speakers (French)

Medium clips, roughly 4–5 minutes of audio:

- English, one speaker (~4 minutes)
- English, two speakers (~5 minutes)

Long-form content, roughly 40 minutes of audio (**preprocessing** was done on the audio before sending to account for file limits):

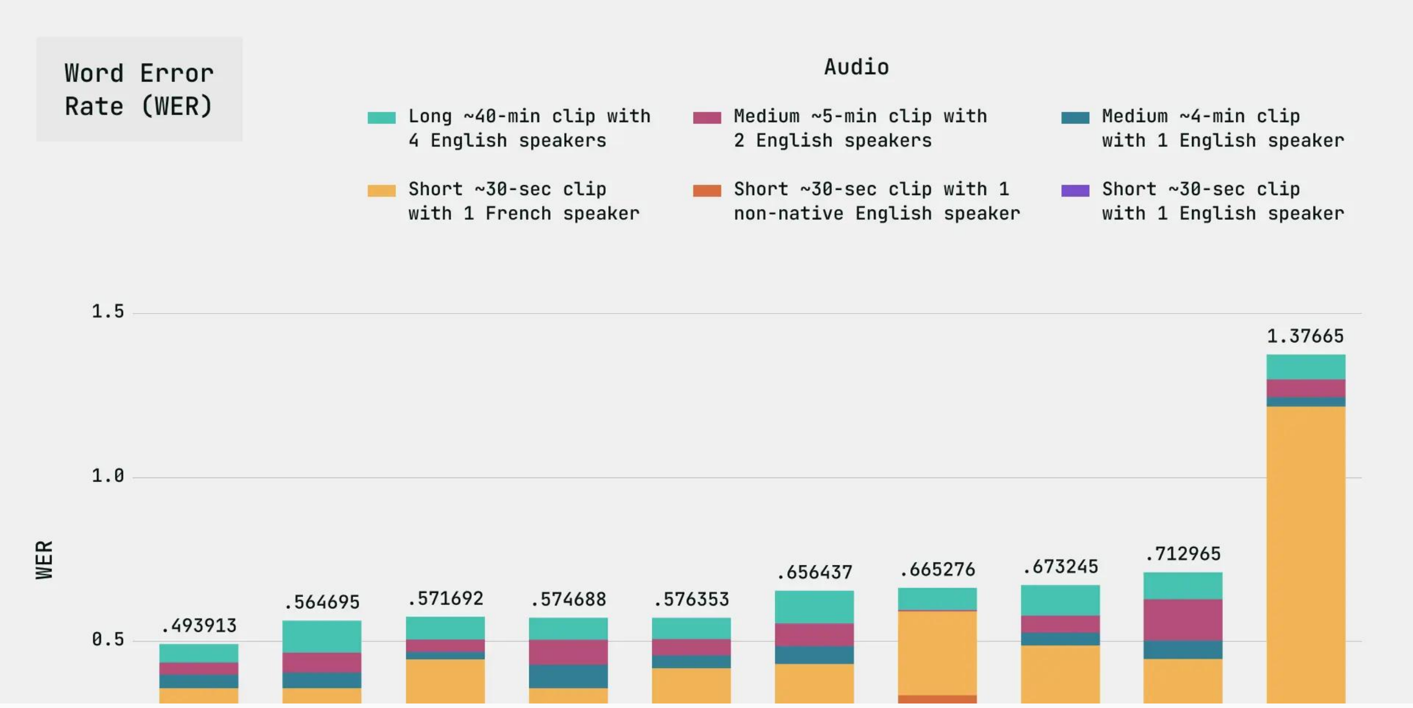
- Panel discussions featuring multiple speakers

Audio clips included a small sample from the **Facebook Voxpopuli dataset**, a large collection of validated transcripts across 18 languages and non-native speech data for short audio. Though most of our sample clips were English content, we included clips with various accents to simulate real-world conversations. We utilized medium-length content from our very own **WillowTree YouTube channel**, and long-form audio came from a Microsoft Research Forum panel discussion.

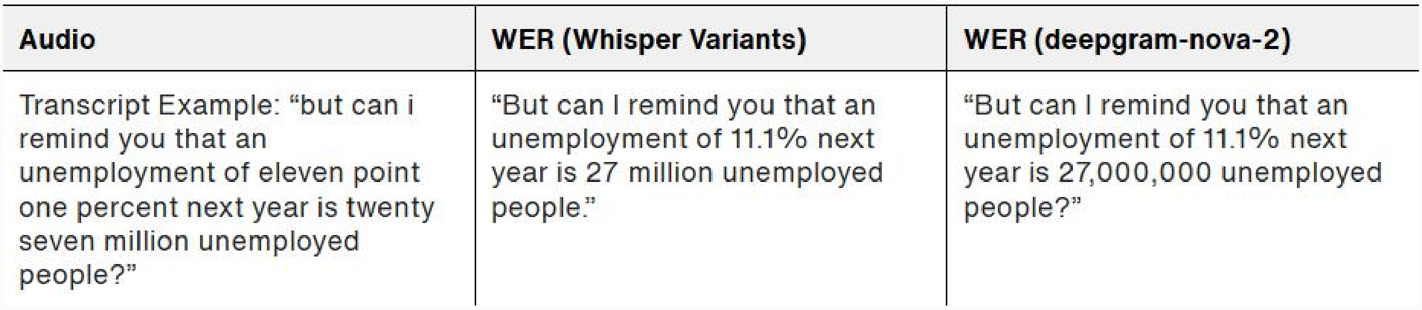
We scored each speech-to-text model based on the metrics WER and WPM, as described earlier. We calculated WER using the jiwer python library using wer_standardize with `RemovePunctuation` enabled. As for WPM, we took the words in the transcript divided by the latency. Last, we evaluated whisper-large-v3-local on an Apple MacBook Pro running a M3 Max chip, 36 GB of memory, and MacOS Sequoia 15.1.

Word error rate (lower is better)

Overall, assemblyai-universal-2 appeared to be the best speech-to-text model we tested. It performed the most consistently across our scenarios, exhibiting the lowest cumulative WER score as shown in the graphic below.



However, it's worth noting that our clips contained instances of spoken numbers, which were transcribed differently by different models. This caused some deviation from the reference transcriptions and may have introduced some inconsistency with the WER calculation. The excerpt below illustrates this.

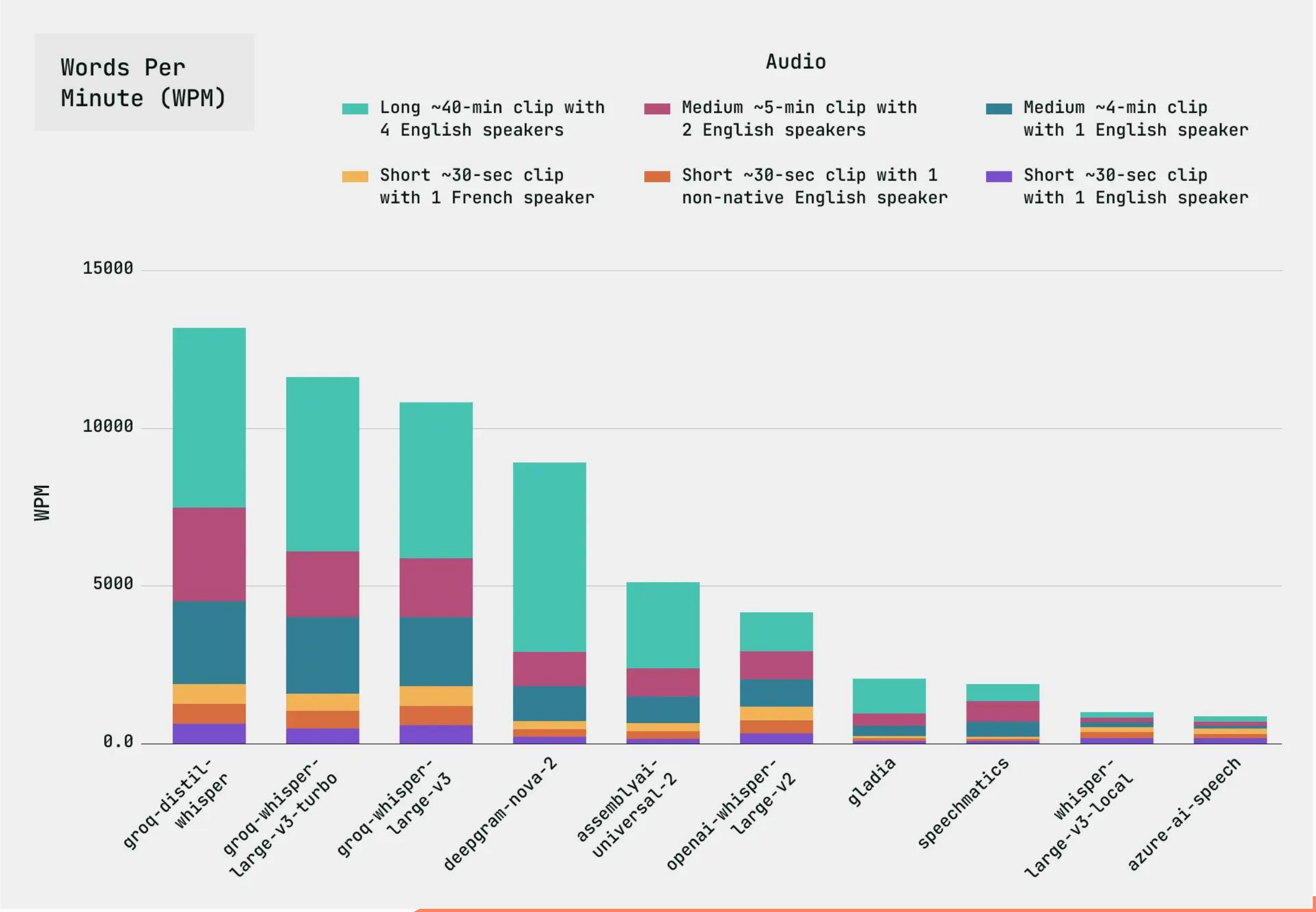


Audio	WER (Whisper Variants)	WER (deepgram-nova-2)
Transcript Example: "but can i remind you that an unemployment of eleven point one percent next year is twenty seven million unemployed people?"	"But can I remind you that an unemployment of 11.1% next year is 27 million unemployed people."	"But can I remind you that an unemployment of 11.1% next year is 27,000,000 unemployed people?"

So, whereas the human transcriber of our reference dataset chose to write out "eleven point one percent next year is twenty seven million," most SST model training solves this problem differently. As we see above, "11.1% next year is 27 million" and "11.1% next year is 27,000,000" are both correct interpretations of the audio, but they're technically wrong because they don't match the source data. In reality, all the SST models performed accurately here.

Words per minute (higher is better)

In terms of transcribing audio the fastest, groq-distil-whisper was the best speech-to-text model we tested. It handled all of our different audio durations for transcription, but with one important caveat: groq-distil-whisper is English only.



Multilingual support

During our tests using a short French clip about 30 seconds long, whisper-large-v3-local performed the best. Outside of running locally, there was a tie between lowest WER across four models:

- assemblyai-universal-2
- speechmatics
- groq-whisper-large-v3
- groq-whisper-large-v3-turbo

Note that all models tested allow for multi-language support (with the exception of groq-distil-whisper), but depending on your use case and language needed, it’s a good idea to check the API documentation to see the specific services available for language support.

If you need help choosing the best speech-to-text models for your apps, we can help you:

- optimize your speech-to-text models for specific tasks (e.g., transcription)
- test different models to find the best application-model fit
- stream content into your app to activate voice-powered experiences