Avalon: ASR for Human–AI Interaction

Jack McIntire
Aqua Voice
San Francisco, CA 94131
jack@withaqua.com

Finnian Brown
Aqua Voice
San Francisco, CA 94131
finn@withaqua.com

1 Introduction

Avalon is a speech recognition model optimized for human–computer interaction. It was trained on a new large-scale audio dataset of real and synthetic human-AI interactions, which improved transcription performance on all tasks and significantly improved performance in domains like software and coding.

2 Background

Today, the best performing speech recognition systems use large-scale unsupervised pretraining on hundreds of thousands or millions of hours of audio. This technique was pioneered by wav2vec 2.0 [Baevski et al., 2020] and was pushed further by Whisper [Radford et al., 2022], which was trained on over 680,000 hours of audio for v1, and later 5,000,000 hours.

Scaling the number of hours significantly improved the performance of speech recognition systems; however, the distribution of pretraining data has led to several real-world usability problems for Whisper and Whisper-derived models.

Much of the publicly available human-labeled audio falls into the following categories:

- Audiobooks (e.g., LibriSpeech)
- Phone calls / meetings
- Television / broadcast news
- Court / parliamentary proceedings

We believe Whisper and similar models were trained largely on this kind of audio data. This distribution contributes to gaps between public-benchmark performance and real-world performance on common human–AI conversation tasks. For example, models like Whisper Large v3 perform well on archaic terms *"perspicacious"*, but struggle on common technical phrases like *"git checkout dev"*.

Avalon targets these specific weaknesses in Whisper. We wanted a model that:

1. is better on programming/coding terms and AI-specific terms.
2. normalizes unfamiliar terms conservatively (domain-aware normalization).
3. has improved robustness on low-quality or distorted audio.

3 Model Data and Training

Avalon was trained on a diverse dataset that includes subsets of publicly available audio datasets, audio gathered from the internet. A subset of data was generated by opt-in users of the Aqua Voice app.

We constructed a data processing pipeline that includes rigorous filtering to maintain label quality, remove low quality samples, and maintain speaker diversity. We supplemented our real samples with synthetically generated data using a process similar to ContextASR-Bench [Wang and collaborators, 2025].

Avalon was trained for two epochs on a cluster of NVIDIA H100 GPUs.

4 Evaluation

4.1 Standard ASR Evaluations

4.1.1 Datasets

We used the official test splits of seven public ASR datasets [hf, 2023]. These include LibriSpeech [Panayotov et al., 2015], SPGISpeech [O'Neill et al., 2021], GigaSpeech [Chen et al., 2021], Earnings-22 [Del Rio et al., 2022], TED-LIUM 3 [Hernandez et al., 2018], AMI [Carletta and et al., 2005], and VoxPopuli [Wang et al., 2021].

Table 1: ASR datasets appearing on the OpenASR leaderboard.

Dataset	Domain	Lang.	Hours	Test	Released	Labels
LibriSpeech	Audiobooks	EN	960	5	2015	Norm.
SPGISpeech	Finance Meetings	EN	5000	100	2021	Punct.+Case
GigaSpeech	Audiobooks/Podcasts/YT	EN	33005	35	2022	Punct.
Earnings-22	Finance Meetings	EN	119	5	2021	Punct.+Case
TED-LIUM 3	TED talks	EN	452	3	2018	Norm.
AMI	Meetings	EN	78	9	2006	Punct.+Case
VoxPopuli	EU Parliament	16	1800	5	2021	Punct.

Punct. = punctuated; Case = cased; Norm. = normalised. *Hours* is total dataset hours; *Test* (hours) is the official test split used when running OpenASR benchmarks.

4.1.2 Performance

Avalon performs well across all eight test splits, achieving the lowest average word error rate of the models that were tested.

Despite being derived from Whisper, Avalon achieves lower word error rates than Whisper Large v3 on seven of the eight test splits. The largest reductions were on AMI, GigaSpeech, and TED-LIUM. Avalon's state-of-the-art performance on GigaSpeech is of particular note because it is the largest and most diverse dataset and is closer to typical real-world usage. Related Whisper-derived systems like CrisperWhisper focus on accurate timestamps and verbatim transcription [Wagner, 2024].

Table 2 reports WER across standard public datasets. We include Avalon and a set of contemporary systems. Avalon outperforms several leading models, including Whisper Large v3, ElevenLabs Scribe v1, and AssemblyAI Best, on average across a suite of standard tests. References for competitor systems: NVIDIA Canary-1B [nvi, 2024a], Voxtral Mini 3B [vox, 2025], and IBM Granite 8B [Saon et al., 2025].

Table 2: Word Error Rate (WER%) across public benchmarks. Lower is better.

Dataset	Avalon	Nvidia Canary 1B	ElevenLabs Scribe v1	Voxtral Mini 3B	OpenAI Whisper Large v3	AssemblyAI Best
AMI	**11.58**	13.90	14.43	16.30	15.99	15.64
Earnings22	11.38	12.19	12.14	**10.69**	11.11	13.54
GigaSpeech	**9.50**	10.12	9.66	10.24	10.12	**9.50**
LibriSpeech (clean)	1.68	**1.48**	1.79	1.88	1.98	1.74
LibriSpeech (other)	3.28	**2.93**	3.31	4.10	4.58	3.11
SPGISpeech	2.10	2.06	3.30	2.37	2.95	**1.81**
TED-LIUM	**3.02**	3.56	3.17	3.68	3.56	3.43
VoxPopuli	7.33	**5.79**	7.20	7.14	8.56	7.47
Average	**6.23**	6.50	6.88	7.05	7.36	7.03

These tests were run using the open-source OpenASR leaderboard repository [hf, 2023].

4.2 Programming Domain Performance

4.2.1 Dataset

To test performance on coding and AI terms, we constructed AISpeak, an evaluation dataset of clips where the speaker is prompting an AI, and often uses jargon and domain-specific terms. As shown in Table 1, the standard ASR evaluation datasets were released several years ago and do not include terms that have recently entered the lexicon, like "Claude Code" or "MCP" [ant, 2024].

We compile a list of these new terms and domain-specific phrases that are common in prompting AIs, but are not present in standard ASR evaluations. We present three variants of AI Speak: AI Speak-10, AI Speak-50, and AI Speak-500, where progressively more difficult terms and phrases are included as the number increases. We constructed the AISpeak evaluations using data from 2025.

Table 3: AISpeak evaluation set details.

Set	Samples	Hours
AISpeak-10	1,278	4
AISpeak-50	3,793	13
AISpeak-500	9,198	31

4.2.2 Performance

Avalon performed well on AISpeak, achieving both lower word error rate for the test split as well as much higher accuracy on specific terms that were highlighted as important for each clip. On the least challenging evaluation, AISpeak-10, Avalon achieved 97.4% accuracy on highlighted terms, compared to 51.5% for NVIDIA Canary 1B and 65.1% for Whisper Large v3.

Table 4: AISpeak Accuracy on Coding and AI Terms

Set	Avalon	NVIDIA Canary 1B	NVIDIA Canary 1B Flash	Voxtral Mini 3B	ElevenLabs Scribe v1	Whisper Large v3	IBM Granite 8B
AISpeak-10	**97.4**	51.5	56.9	59.5	78.8	65.1	54.7
AISpeak-50	**97.5**	71.8	74.5	79.4	86.7	82.4	72.8
AISpeak-500	**95.8**	74.1	76.1	82.9	87.5	84.9	75.0

The relatively poor performance of NVIDIA's Parakeet family of models on AISpeak is noteworthy, because on standard ASR benchmarks they perform well. On AISpeak they scored between 50-75% and were the least accurate models we tested. We speculatively concluded that the Parakeet models may be overfit to public audio datasets [nvi, 2024b].

Below are selected examples illustrating Avalon's performance on coding-related utterances compared to Whisper Large v3 and other systems.

```
Avalon  : Can you add that to my zshrc?
Whisper : You add that to my C, short C.
Human   : Can you add that to my zshrc?
```

Listing 1: Comparison for "zshrc"

```
Avalon  : Make a fully featured PyTorch alternative.
Whisper : Make a fully featured high torch alternative.
Human   : Make a fully featured PyTorch alternative.
```

Listing 2: Comparison for "PyTorch"

```
Avalon  : Let me make something plain. There's only one instance where a sushi belt should be used in
     Factorio, and that is when you're building a hauler spaceship.
Whisper : Let me make something plain. There's only one instance where a Sushi Belt should be used in
     Factorio: Gamma, and that is when you're building a hull or spaceship
Human   : Let me make something plain. There's only one instance where a sushi belt should be used in
     Factorio, and that is when you're building a hauler spaceship.
```

Listing 3: Comparison for "sushi belt" in Factorio

```
Avalon  : Grok 4 beats GPT-5 on Arc AGI.
Whisper : Brock 4 beats GPT-5 on ARK AGI.
Human   : Grok 4 beats GPT-5 on Arc AGI.
```

Listing 4: Comparison for "Grok 4" vs "Arc AGI"

```
Avalon  : In my Vercel configuration, I was having an issue with node versions.
Whisper : in my like, Versacell configuration, I was having an issue with node versions.
Human   : In my Vercel configuration, I was having an issue with node versions.
```

Listing 5: Comparison for "Vercel" node versions

```
Avalon  : I've tried running this with GPT-4o, GPT-4.1, and o3.
Whisper : I've tried running this with GPT-4.0, GPT-4.1, and GPT-03.
Human   : I've tried running this with gpt-4o, gpt-4.1, and o3.
```

Listing 6: Comparison for complex AI interaction

In this example, Avalon correctly identifies model names and their respective casing and formatting, and it also avoids hallucinating a "GPT" prefix to "o3", which did not appear in the speech.

```
Avalon  : We use uv as our package manager.
Whisper : We use UV as our package manager.
Human   : We use uv as our package manager.
```

Listing 7: Comparison for "uv" vs "UV"

In this example, Avalon correctly normalizes the coding term 'uv' to its lowercase form, matching common usage.

References

Open asr leaderboard. Hugging Face Space + code repository, 2023. URL https://huggingface.co/spaces/hf-audio/open_asr_leaderboard.

Introducing the model context protocol. Anthropic News, 2024. URL https://www.anthropic.com/news/model-context-protocol.

Canary-1b (nvidia nemo). Hugging Face model card, 2024a. URL https://huggingface.co/nvidia/canary-1b.

Pushing the boundaries of speech recognition with nvidia nemo parakeet asr models. NVIDIA Developer Blog, 2024b. URL https://developer.nvidia.com/blog/pushing-the-boundaries-of-speech-recognition-with-nemo-parakeet-asr-models/.

Voxtral mini 3b. Hugging Face announcement/blog, 2025. URL https://huggingface.co/blog/voxtrol-mini-3b.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, 2020. URL https://arxiv.org/abs/2006.11477.

Jean Carletta and et al. The ami meeting corpus: A pre-announcement. In *MLMI*, 2005.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, and et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *INTERSPEECH*, 2021. URL https://arxiv.org/abs/2106.06909.

Miguel Del Rio, Peter Ha, Quinten McNamara, and et al. Earnings-22: A practical benchmark for accents in the wild. *arXiv:2203.15591*, 2022. URL https://arxiv.org/abs/2203.15591.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. *arXiv:1805.04699*, 2018.

Patrick K. O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, and et al. Spgispeech: 5000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. In *INTERSPEECH*, 2021. URL https://www.isca-archive.org/interspeech_2021/oneill21_interspeech.html.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *ICASSP*, 2015.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv:2212.04356*, 2022. URL https://arxiv.org/abs/2212.04356.

George Saon, Avihu Dekel, Alexander Brooks, and et al. Granite-speech: open-source speech-aware llms with strong english asr capabilities. *arXiv:2505.08699*, 2025.

Steve Wagner. Crisperwhisper. GitHub repository, 2024. URL https://github.com/stevewagner/cwhisper.

Changhan Wang, Morgane Rivière, Ann Lee, and et al. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv:2101.00390*, 2021.

Y. Wang and collaborators. Contextasr-bench. GitHub repository, 2025. URL https://github.com/ctcdecode/ContextASR-Bench.