Received: 27 September 2022 | Accepted: 31 October 2022

DOI: 10.1002/aaai.12071

COLUMN

Is AI at human parity yet? A case study on speech recognition

Ian Beaver

Verint Systems Inc, Melville, New York, USA

Correspondence
Ian Beaver, Verint Systems Inc, 175 Broadhollow Rd, Ste 100, Melville, NY, USA.
Email: ian.beaver@verint.com

Abstract

Claims have been made that speech recognition has achieved human parity, yet this does not appear to be the case in the real-world applications that rely on it, especially for non-native speakers. This then begs the questions: What does it even mean for an AI system to reach human parity? How is progress towards that goal being measured? This article focuses on the current state of speech recognition and the recent developments in benchmarking and measuring performance of AI models built for speech processing. Through the shift away from single metric benchmarks and specialized models and towards evaluating collections of diverse challenging tasks and generalized models, the ultimate goal of true human parity in commercial speech processing applications is hopefully on the near horizon.

Like many people, I work remotely for a company with employees all over the world. This work environment requires daily video calls with coworkers, many of whom are non-native English speakers (in the field of linguistics, native speakers of a language are referred to as L1 and non-native as L2). Frequently, we use automatic transcription to take notes on what was discussed during the call. These call transcripts range from very good to unusable depending on who is talking, terminology, and various environmental factors. While this behavior is not surprising to people who have worked on or with automatic speech recognition (ASR) engines, others have been surprised at the pervasive errors given the many recent advancements in the field and some highly publicized claims of human parity in the task.

It is true that many breakthroughs have happened in the speech arena over the last decade, and there are many fields that depend on good quality speech recognition such as conversational AI, smart speakers, and autonomous vehicles; all of which are continuing to push research in speech recognition forward. This intensive research focus combined with better algorithms, data availability, and better hardware has led to a steady increase in ASR performance. Performance in speech recognition, as well as nearly all human language tasks, is measured against the human performance on a particular set of examples that are organized into a benchmark corpus. The ability for a system to meet or exceed the measurement of human performance is referred to as human parity (in the tested corpus).

For ASR, this milestone was first claimed in a 2016 research paper by Microsoft (Xiong et al., 2016) reporting that for the first time, they have achieved human parity in word error rate[i] (WER) on the Switchboard benchmark (5.8% WER) while also achieving 11% WER on the Call-Home benchmark, which is known to be more challenging to transcribe. In addition, the reported decoding speed was only 1.38× real time, which is in the realm of usability for some commercial systems. This announcement was highly publicized even in mainstream media outlets[ii]. A follow-up paper in 2017 claimed further improvement to 5.1% WER on Switchboard but with no report on decoding speed (Xiong et al., 2018). Also in 2017, Google announced a 4.9% WER (on some undisclosed benchmark) at its annual I/O developer conference[iii]. As Switchboard and CallHome are both corpora of human-to-human telephone conversations, a reader of such highly publicized announcements may conclude that an ASR system that meets human parity in these benchmarks should be able to transcribe speech as good as a human *in general*. Yet with all this advancement in the field, user experiences with real-world speech systems do not yet reflect true human parity.

Looking beyond the Switchboard and CallHome benchmarks, speech WER was measured to be 78–89% for individuals with self-reported disordered speech (Green et al., 2021) and 16–23% on those with "no abnormalities" (De Russis and Corno, 2019). In 2020, comparison of five major speech recognition services reported 14–18% WER on transcription of podcasts with L1 speakers[iv]. A more recent comparison of cloud vendors from June 2022 using YouTube clips reported 10–14% WER[v]. Life is harder for L2 speakers as we still see on average 32–59% WER across cloud service speech platforms (Cumbal et al., 2021). To run an experiment yourself, connect to a video conferencing call with several L2 speakers discussing some business or technical topics, turn on a live transcription service of your choosing, and compare its performance to your ability to understand the speakers.

Therefore, the claims of "human parity" marketed in 2016 still stand as an uncashed check for many current users of speech recognition systems (unless you happen to be an adult native US English speaker, without a speech disorder, with a good microphone, in an ideal speaking environment, and discussing general vocabulary topics). In defense of Microsoft, the speech group openly acknowledged parity in a single benchmark does not equal parity in all scenarios in presentations of their paper[vi], and this clarification gets to the core of this article: beating humans in a benchmark language task does not mean an AI system has reached human parity in the general sense. So how can the research community better quantify how far away they are from this ultimate goal of human performance they are reaching for? What trends have surfaced in the last few years that will help speech recognition finally meet human parity in all but the hardest of situations where even humans have difficulty?

Speech recognition is hard for many reasons beyond accents and background noise. Training and evaluation data do not always accurately reflect human performance as quick versus thorough annotations lead to big WER differences (Glenn et al., 2010), benchmarks and training data do not always contain good population representations (L1 vs. all variations of L2), and audio can be captured in a variety of means (smart speaker in a large room, laptop microphone in a crowded airport, cell phone microphone in a hallway that echoes, etc.). There are not always public benchmark datasets available in the native languages of the target users or the topics discussed in benchmark corpora are missing a lot of vocabulary that would be important for a practitioner in a specific field to measure performance on (e.g., medical or technical jargon).

In addition, in the speech processing community, it has been widely discussed why WER is not a perfect metric to measure the quality of interaction with a speech application as the metric assumes all words have equal value, and variations of equivalent tokens may not be considered equal by the scoring script (e.g., Apt. versus Apartment). Yet there has not been a superior metric adopted to date, so WER still appears the primary measure of success in ASR literature. One of the issues with WER is it does not capture errors in the meaning of spoken language. In downstream applications using ASR output, even if a system has a WER of 1% but those are the most critical 1% of words for the task to succeed the user will still have trouble completing their task. Conversely, a system may have a 15% WER and users are able to succeed at a task just fine because either the words with errors are not critical to the task (e.g., substituting "uh" for "a") or downstream systems are detecting and correcting for the ASR errors before applying various post processing, like is common in smart speakers or conversational AI (Ponnusamy et al., 2020). In this case, users may have a good experience despite the ASR component performing well below general human parity.

Therefore, when quantifying human parity, there are two primary issues to be considered with the construction of human language benchmarks in general: how closely does the benchmark reflect the real-world environment the system is expected to perform in and does the evaluation metric measure the actual effect of system errors on its users? When a system excels on a benchmark and/or metric that is lacking these qualities and the results are widely publicized, there are nothing but unmet expectations to be gained when such systems are released to users. Human parity in a language task as general as speech recognition is very hard to represent in a single benchmark and even harder to evaluate for any given system application outside of a case study directly with the intended users.

To combat this challenge, recent work has been done in many speech and language processing fields to create large aggregate benchmarks that better capture how systems perform across a variety of loosely related tasks. In text processing, there was the creation of the GLUE (Wang et al., 2018) followed by SuperGLUE (Wang et al., 2019) benchmarks. The latter is a collection of 10 challenging tasks covering various aspects of language processing such as reading comprehension, textual entailment, and answering yes/no questions about a passage of text. In order to excel in the SuperGLUE benchmark, a model needs to learn more than just grammar and syntax or memorize training data but must now work to understand language usage in context and implication. Following this paradigm, the BIG-bench benchmark (Srivastava et al., 2022) was created as a collaborative and living benchmark aimed at collecting as much variety and coverage of language tasks as possible to more effectively probe the capabilities of ever-growing large language models. It contains 211 tasks by over 442 authors and 132 different institutions at the time of this writing.

While not as expansive in terms of task evaluations as those available in text processing, to provide a more robust measure of speech processing performance, the Speech processing Universal PERformance Benchmark (SUPERB) was released in 2021 containing 10 tasks such as speaker identification, keyword spotting, speaker diarization (separating speakers in a single audio stream), and speech recognition (Yang et al., 2021). This benchmark was extended by SUPERB-SG in 2022 with increased diversity and difficulty of tasks such as speech translation, voice conversion (convert speech from an arbitrary speaker into a target speaker such as a celebrity), and speech enhancement (Tsai et al., 2022). While some of these tasks are hard to measure human performance in, after all not many people can convincingly imitate any given target speaker, to do well at these diverse tasks helps force models to excel at speech processing in general, which is the ultimate goal for AI. The speech processing community has also seen some large corpora released recently to help coverage quality in high-resource languages such as Libri-light (Kahn et al., 2020), GigaSpeech (Chen et al., 2021), and WenetSpeech (Zhang et al., 2022). The public availability of more large varieties of speakers in diverse environments will also help models to improve for the countless speaker/environment combinations seen in the real world.

This recent pivot in the speech processing community from using a few gold-standard benchmarks targeting a single metric to evaluating a panel of very challenging tasks with different metrics where models must generalize in order to perform well will hopefully go far to advance the quality of commercial ASR systems. After all, in a short period of time, the GLUE and SuperGLUE benchmarks challenged the language processing community to reduce focus on highly specialized task-specific language model construction and work on more generalized language understanding models that can easily adapt to a variety of tasks. With the increased availability of such large and diverse training datasets and more diverse benchmarks as well as the recent application of self-supervised learning to take advantage of large unlabeled audio corpora (Chen et al., 2022), we should finally see speech recognition approach true human parity, at least in high resource languages. As for L2 speakers and their struggle to be understood by AI systems, hopefully, this future is not far away.

CONFLICT OF INTEREST

The author declares that there is no conflict.

ORCID

Ian Beaver https://orcid.org/0000-0003-0865-1214

ENDNOTES

[i] https://en.wikipedia.org/wiki/Word_error_rate

[ii] For example: https://www.newsweek.com/microsoft-speech-recognition-achieves-human-parity-511538

[iii] https://venturebeat.com/business/googles-speech-recognition-technology-now-has-a-4-9-word-error-rate/

[iv] https://www.rev.com/blog/speech-to-text-technology/the-podcast-challenge-testing-rev-ais-speech-recognition-accuracy

[v] https://www.voicegain.ai/post/speech-to-text-accuracy-benchmark-june-2022

[vi] https://www.microsoft.com/en-us/research/wp-content/uploads/2017/01/HumanParity.pdf

REFERENCES

Chen, G., S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, et al. 2021. “Gigaspeech: An Evolving, Multi-domain ASR Corpus with 10,000 Hours of Transcribed Audio.” *arXiv preprint arXiv:2106.06909*.

Chen, S., C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, et al. 2022. “WavLM: Large-scale self-supervised pre-training for full stack speech processing.” *IEEE Journal of Selected Topics in Signal Processing* 16(6): 1505–18.

Cumbal, R., B. Moell, J. D. Aťguas Lopes, and O. Engwall. 2021. ““You don’t understand me!”: comparing ASR results for L1 and L2 speakers of Swedish.” In *Proceedings of the Interspeech 2021*.

De Russis, L., and F. Corno. 2019. “On the impact of dysarthric speech on contemporary ASR cloud platforms.” *Journal of Reliable Intelligent Environments* 5(3): 163–72.

Glenn, M., S. Strassel, H. Lee, K. Maeda, R. Zakhary, and X. Li. 2010. “Transcription methods for consistency, volume and efficiency.” In *Proceedings of the Seventh Interna- tional Conference on Language Resources and Evaluation (LREC’10)*.

Green, J. R., R. L. MacDonald, P. P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, et al. 2021. Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases. In Interspeech (pp. 4778–82).

Kahn, J., M. Rivie‘re, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazareť, J. Karadayi, et al. 2020. “Libri-light: a benchmark for asr with limited or no supervision.” In *Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7669–73. IEEE.

Ponnusamy, P., A. R. Ghias, C. Guo, and R. Sarikaya. 2020. “Feedback-based self-learning in large-scale conversational AI agents." In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 13180–7.

Srivastava, A., A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, et al. 2022. "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models." *arXiv preprint arXiv:2206.04615*.

Tsai, H.-S., H.-J. Chang, W.-C. Huang, Z. Huang, K. Lakhotia, S.-W. Yang, S. Dong, et al. 2022. "SUPERB-SG: Enhanced Speech Processing Universal Performance Benchmark for Semantic and Generative Capabilities." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8479–8492.

Wang, A., Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2019. "SuperGLUE: a stickier benchmark for general-purpose language understanding systems." In *Proceedings of the* Advances in *Neural Information Processing Systems*, 32.

Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2018. "GLUE: A Multi-task Benchmark and Analysis Platform for Natural Language Understanding."

Xiong, W., J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. 2016. "Achieving Human Parity in Conversational Speech Recognition." *arXiv preprint arXiv:1610.05256*.

Xiong, W., L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke. 2018. "The Microsoft 2017 conversational speech recognition system." In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5934–8. IEEE.

Yang, S.-W., P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, et al. 2021. "Superb: Speech Processing Universal Performance Benchmark." *arXiv preprint arXiv:2105.01051*.

Zhang, B., H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, et al. 2022. "WenetSpeech: a 10000+ hours multi-domain mandarin corpus for speech recognition." In *Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6182–6. IEEE.

AUTHOR BIOGRAPHY

Ian Beaver (PhD, University of New Mexico) has worked on topics surrounding human–computer interactions such as gesture recognition, user preference learning, and communication with multimodal intelligent virtual assistants since 2005. He has authored over 40 patents within the field of human language technology and regularly serves as a PC member in many top AI and NLP conferences. Ian is a Chief Scientist at Verint Systems Inc where he works to optimize human productivity in contact centers by way of automation and augmentation and improve customer self-service experiences through the application of conversational AI.

How to cite this article: Beaver, I. "Is AI at human parity yet? A case study on speech recognition." *AI Magazine* 2022.43: 386–389. https://doi.org/10.1002/aaai.12071