Ursa: Benefits of Scaling Self-Supervised learning for Automatic Speech Recognition

Bethan Thomas Ana Olssen John Hughes Benedetta Cevoli Jamie Dougherty

Speechmatics

Summary of content

- **Scaled self-supervised learning**: We describe our Ursa project where we scaled our SSL model from 500M to 2B parameters and trained with over 1 million hours of audio from approximately 50 languages.
- **ASR performance gains**: We demonstrate an overall 22% improvement in accuracy compared to our previous model through scaling SSL on English, and gains on a wide variety of lower resource languages.
- **Increased diversity**: We achieve industry-leading performance across a range of diverse voices.
- **Improved sample efficiency**: We also show that our scaled SSL model can outperform a smaller SSL model with 300x less labelled data.

Introduction

We first train a **self-supervised learning (SSL) model**. This uses an efficient transformer variant that learns rich acoustic representations of speech from unlabeled data.

We then use paired audio-transcript data in a second stage to train an **acoustic model** that learns to map self-supervised representations to phoneme probabilities.

The predicted phonemes are then mapped into a transcript by using a **language model** to identify the most likely sequence of words.

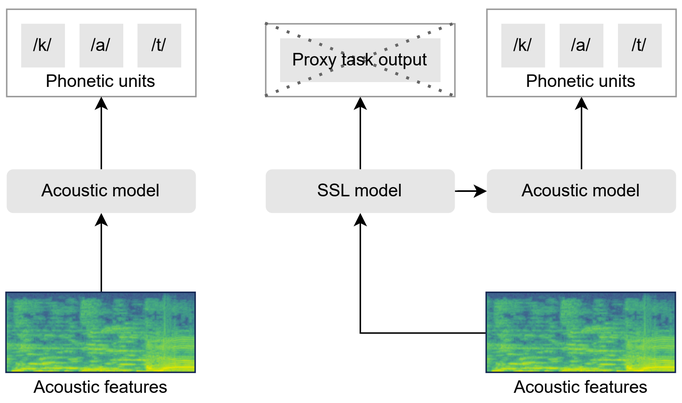

Figure 1. Architecture of standard ASR contrasted with SSL based ASR

Scaled self-supervised learning

We found that the scaling behaviour of acoustic SSL models is similar to that of large language models. Therefore we scaled our model both in terms of size and training data.

- **Model size**: We scaled our SSL model to 2 billion parameters.
- **Training data**: When model size is increased, the capacity of the model to learn from increased training data is also increased. Therefore we also scaled our SSL training data to over 1 million hours spanning around 50 languages.

We evaluated our system across a range of publicly available testsets by calculating word error rate (WER). We compare to our previous model which used less powerful SSL. We also compare to Whisper [7] which is a publicly available model which does not use SSL.

	Ursa	Baseline	Whisper
Weighted average	11.96	15.36	15.95

Table 1. Averaged results on a wide range of publicly available testsets. Ursa is our scaled SSL model, baseline is our previous best SSL model, Whisper is a non-SSL model.

Sample efficiency

SSL models produce rich representations of audio which enable easier learning of the mapping between speech and text. The more powerful SSL models learn stronger representations which should lead to better ASR performance and increased sample efficiency. Greater sample efficiency means less required labeled ASR training data, and also quicker training times.

We tested sample efficiency on English by comparing 2 sizes of SSL model and training on increasingly limited labeled data.

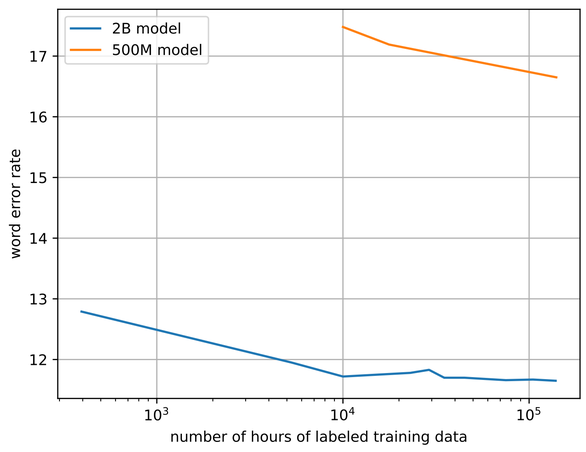

Figure 2. Results from 2 SSL models with different parameter sizes, trained on differing amounts of labeled data.

- Progress saturates quickly above 10,000 hours with our more powerful SSL model.
- For the low-resource regime, performance is still very strong.
- Scaling SSL leads to greater sample efficiency and generally better performance.

Diversity

By scaling to 2 billion parameters, our models are now capable of learning richer acoustic features from unlabeled multi-lingual data, allowing us to understand a larger spectrum of voice cohorts.

Speech-to-text systems have been shown to exhibit systematic inaccuracies or biases towards groups of speakers with varying age, gender, and other demographic factors [8, 4, 6]. Artificial intelligence bias in speech-to-text not only affects the reliability of speech technologies in real-world applications but it can perpetuate discrimination at a large scale.

We evaluated our model using WER on different English accents using the Common Voice dataset [1] and on different demographics using the CORAAL [3] and Casual Conversations [5] datasets.

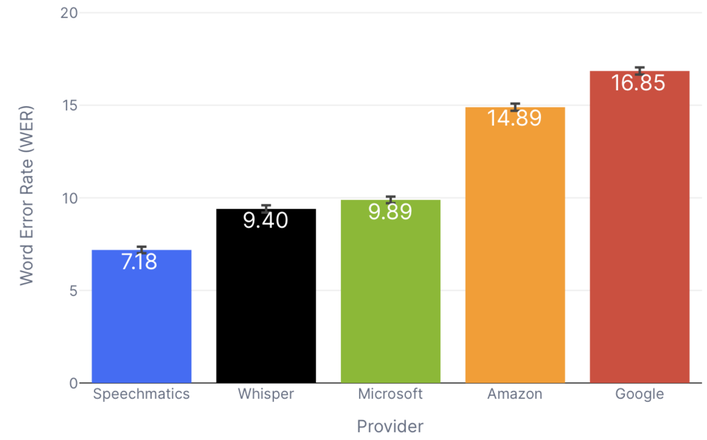

Figure 3. WER of different commercial ASR systems for different English accents based on the Common Voice dataset

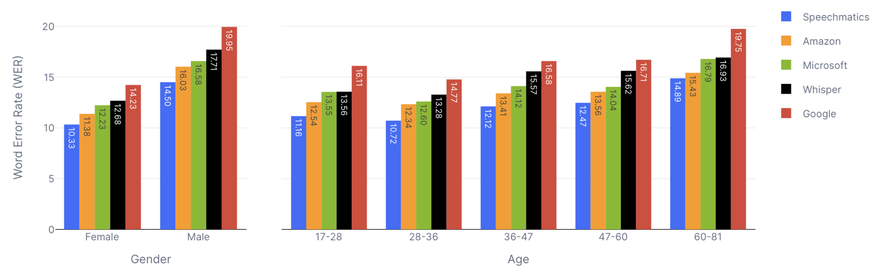

Figure 4. WER of different commercial ASR systems on combined CORAAL and Casual Conversations datasets broken down by gender and age

Language coverage

Scaling traditional ASR models relies on increasing the amount of labeled training data. This is not possible for many of the world's languages. However, SSL relies on unlabeled data which is more readily available. We show that we are able to achieve excellent results on a range of languages spanning 1000s to 10s of hours of labeled training data. This is more evidence of the power of scaled SSL.

We evaluated performance on the FLEURS dataset [2] across 41 languages, and again compared to our previous best model and Whisper.

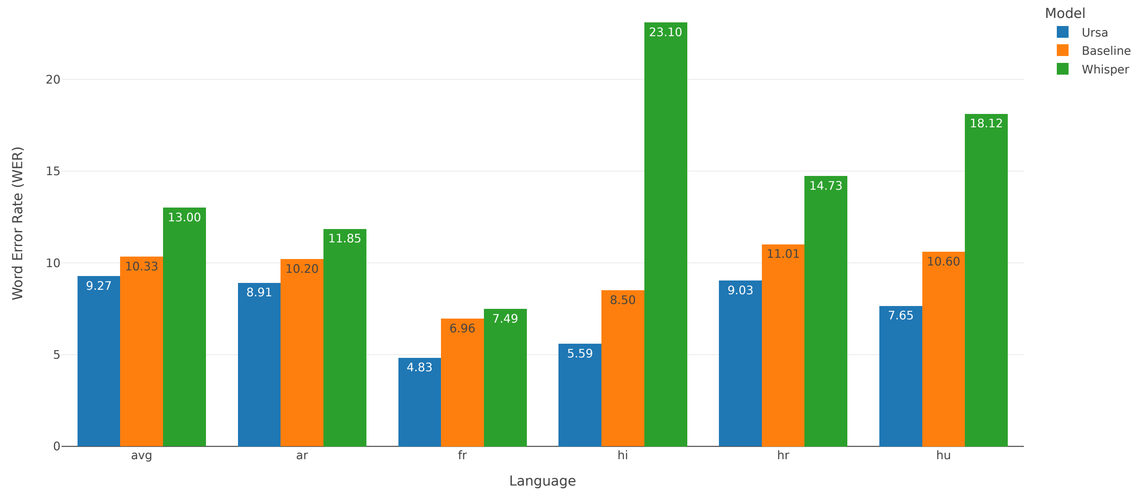

Figure 5. Results on FLEURS testset averaged across 41 languages, and individual language results on Arabic, French, Hindi, Croatian and Hungarian. Ursa is our scaled SSL model, baseline is our previous best SSL model, Whisper is a non-SSL model.

References

[1] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

[2] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE, 2023.

[3] Tyler Kendall and Charlie Farrington. The corpus of regional african american language. version 2021.07. eugene, or: The online resources for african american language project, 2021.

[4] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[5] Chunxi Liu, Michael Picheny, Leda Sarı, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. Towards measuring fairness in speech recognition: casual conversations dataset transcriptions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6162–6166. IEEE, 2022.

[6] Joshua L Martin and Kelly Elizabeth Wright. Bias in automatic speech recognition: The case of african american language. *Applied Linguistics*, 2022.

[7] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

[8] Rachael Tatman. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 53–59, 2017.