Speechmatics

Self-Supervised Learning:

A Step Closer to Autonomous Speech Recognition

November 2021

WHITEPAPER

Speechmatics.com

Say Hello to the Most Powerful, Inclusive and Accurate Speech Recognition on the Planet

At Speechmatics, our mission is to understand every voice. We want to reshape speech recognition, so it doesn't just work for some, but for everyone.

To do this we believe it's not simply about incremental changes. To truly make positive changes in the industry takes bold, forward-thinking action and advances. We set out to rebuild our machine learning from the ground up, looking at models that can learn from all voices, that require minimal human intervention and that can adapt as language evolves.

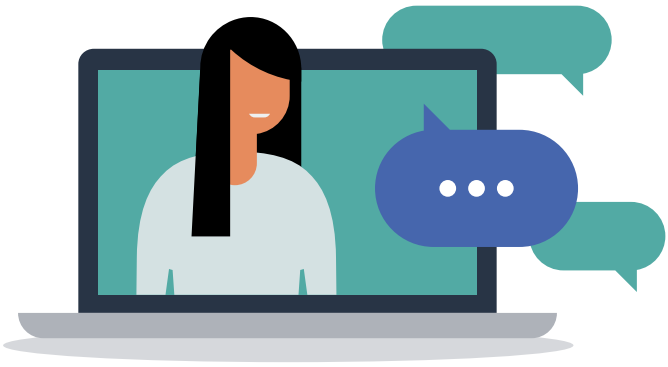
After twelve months of research, innovation and rebuilding, we're ready to introduce to the world Autonomous Speech Recognition (ASR).

The Present and The Future

Traditional speech recognition – also known as 'Automatic' Speech Recognition – has always relied on consuming large quantities of data from a narrow set of speakers. This process has always been achieved through the use of human labeled data, such as audiobooks, where one speaker reads an entire book.

'Autonomous' Speech Recognition, on the other hand, delivers a step-change in accuracy and inclusion by leveraging a wide range of voices using the scale and diversity of the internet. The first step on this journey is introducing and understanding self-supervised learning, the machine learning capability which we've introduced into our training.

We're here to talk through that change and what it means for the future of speech recognition.



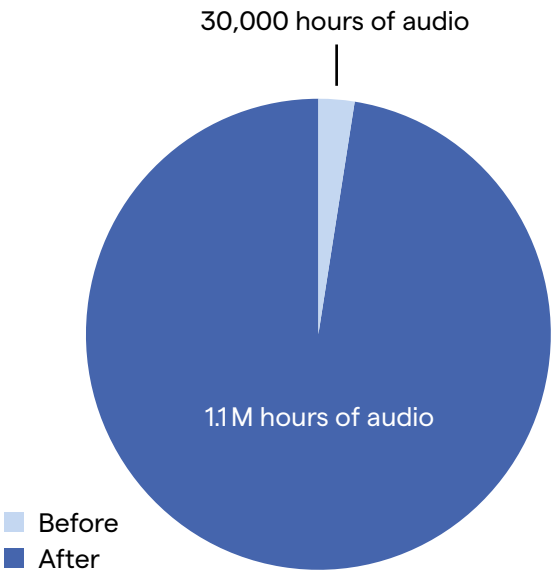
More Voices. Meaningful Change.

To establish meaningful change in machine learning, we've always relied on three areas: improved algorithms, increased computing power and, lastly, the amount of data available to the system.

For too long the biggest obstacle in commercial speech recognition has been the availability of labeled data. While it's true, labeled data in some areas (particularly clear, well-edited American English) have been plentiful, other languages, demographics, and lesser-quality recordings have been harder to come by. This has led to some voices being left behind.

No matter what our speech recognition is ultimately used for, be it broadcast subtitling, reviewing calls in contact centers or a host of other real-world applications, this shortage of labeled data has caused a bottleneck in progress.

We knew if we truly wanted to be able to understand every voice, our approach to data would have to change.



Audio trained on (hours) before vs after unlocking self-supervised learning.

Unlabeled Data: Removing the Middleman

Producing labeled data takes time and money. In speech-to-text, for every audio file with an annotated transcript there's a human listening, pausing after a few words and manually inputting the words spoken.

What's perhaps most surprising about this process is just how inaccurate humans can be. Research by The Linguistics Data Consortium found that in "very careful multiple transcriptions" the Word Error Rate was between 4.1% and 4.5%. When asked to deliver "quick transcription" these levels went as high as 9.6%, meaning a word in every ten, was a mistake.

Working with unlabeled data removes this need for human intervention. The datasets therefore become exponentially greater, with every broadcast, every podcast, every YouTube video now something that offers a wealth of unlabeled data to train on.

To illustrate just how large in scale this change is, before we unlocked self-supervised learning, we were training on around 30,000 hours of audio content. Now, that number is closer to 1,100,000 hours of audio.

> **The Word Error Rating (WER) decreased significantly. Increased accuracy, means more happy customers, and higher margins.**
>
> **HappyScribe**

Advanced Algorithms. Unlocking Data.

In the past, supervised learning from labeled data was the only way to ensure the levels of accuracy needed to feel confident the text was truly representative of the spoken words. We know this approach is costly and narrow and we know it still doesn't guarantee us perfect levels of accuracy – especially when done at pace.

When trained with self-supervised learning, our models can autonomously learn to spot salient patterns in unlabeled data. These models can offer deep and meaningful representations of human language and speech, all without the need for human supervision.

This means Speechmatics ASR can learn by using wider datasets from a huge variety of sources that include both labeled and unlabeled data direct from the internet as well as specialist sources.

It's our goal to ultimately progress to Continuous Learning, meaning we'll learn from new data as it appears online and is ingested into our database. So, for example, when words like COVID-19 enter the lexicon, they'll automatically be updated without the need for human input.

By leading the industry from Automatic Speech Recognition towards Autonomous Speech Recognition we're using the latest techniques in machine learning, including the introduction of self-supervised models, to take things to the next level.

The result? The models are the most accurate we've ever created.

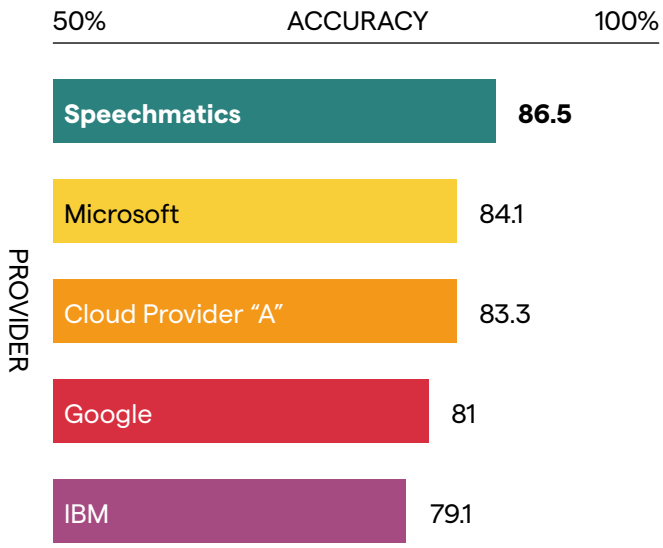
> **Autonomous Speech Recognition has given us the largest step change in accuracy I've seen in my career.**
>
> **Will Williams**
> VP Machine Learning at Speechmatics

> **Our second set of evaluations, focusing on further content types, were every bit as impressive as the first. Improvements across the board!**

Lee Worth
ASR & Live Captioning Operational Excellence Lead, Red Bee Media

Accuracy is Everything

To show how accurate our results are, we needed to assess our competitors. Based on an internal test set of six hours of representative audio (English language audio taken from: meetings, earning calls, journalism, online videos and more), we've seen a huge leap in our accuracy. A leap that puts us well above our competitors.

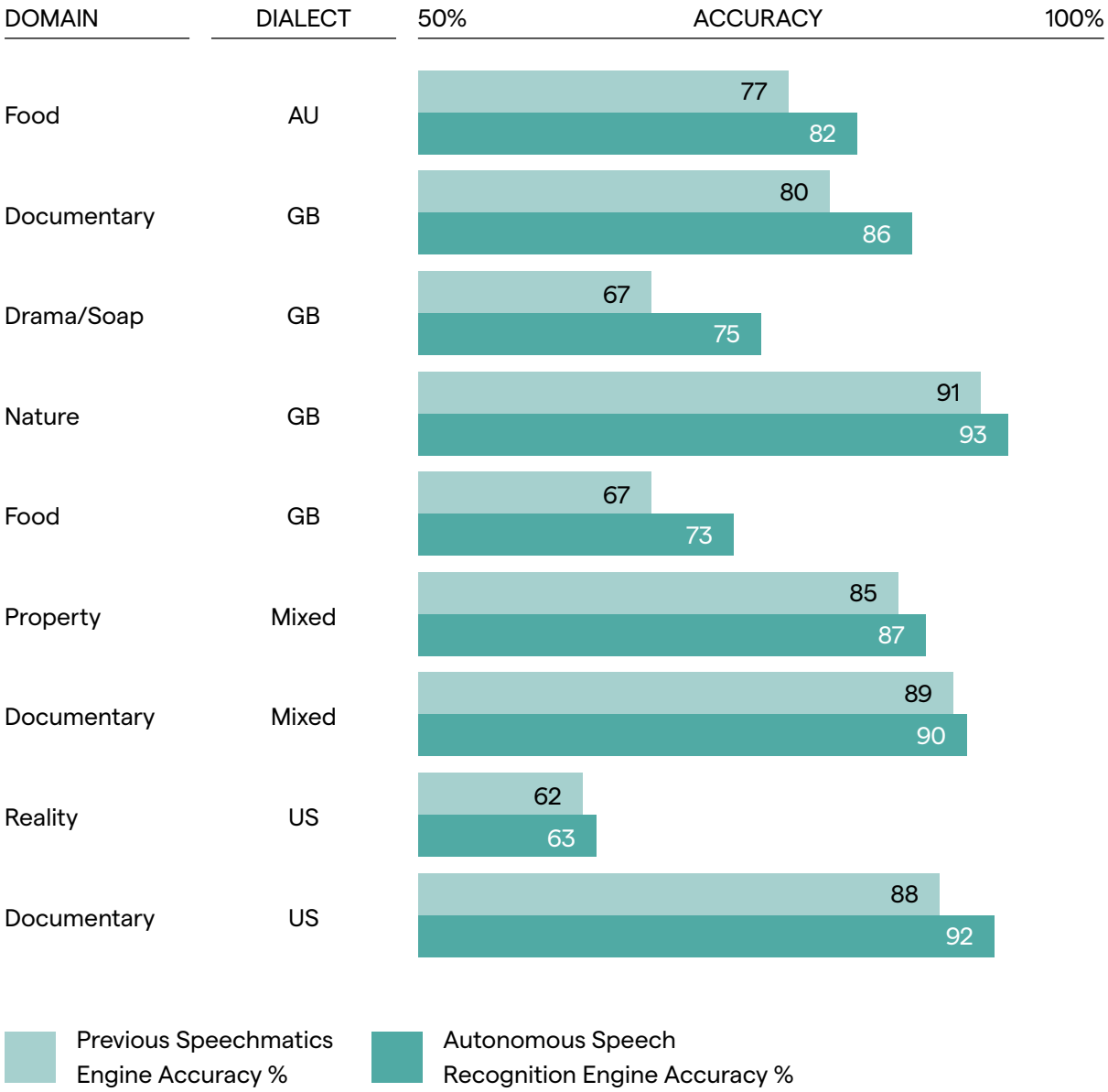


> **Speechmatics continues to prioritize the accuracy of their long-form, content-independent ASR engine. This is an essential component in delivering the highest quality captions to our customers.**

Roger Zimmerman
Chief of R&D at 3Play Media

Leading the way in speech recognition (and being above our competitors for accuracy) has always been a standard for Speechmatics, but with the new self-supervised models the results push the gap wider than ever.

Independent research from Red Bee Media (below) – leading providers of broadcast media captioning – found that when comparing our previous engine to our new Autonomous Speech Recognition there was another extremely positive uptick in accuracy.



In the world of the TV documentary, we were getting accuracy percentages over ninety, with increases of over eight percent in TV Dramas and Soaps. While Reality TV still causes the same challenges it always has – with cross talk and noisy backgrounds – we're seeing an improvement there too. One which, we believe, will only get better with Autonomous Speech Recognition.

While it's always good to see our accuracy ratings topping tables against our competitors, what's also exciting us, is that our team is making the gap between where we were before and where we are now, even wider.

And we firmly believe, this is just the beginning.

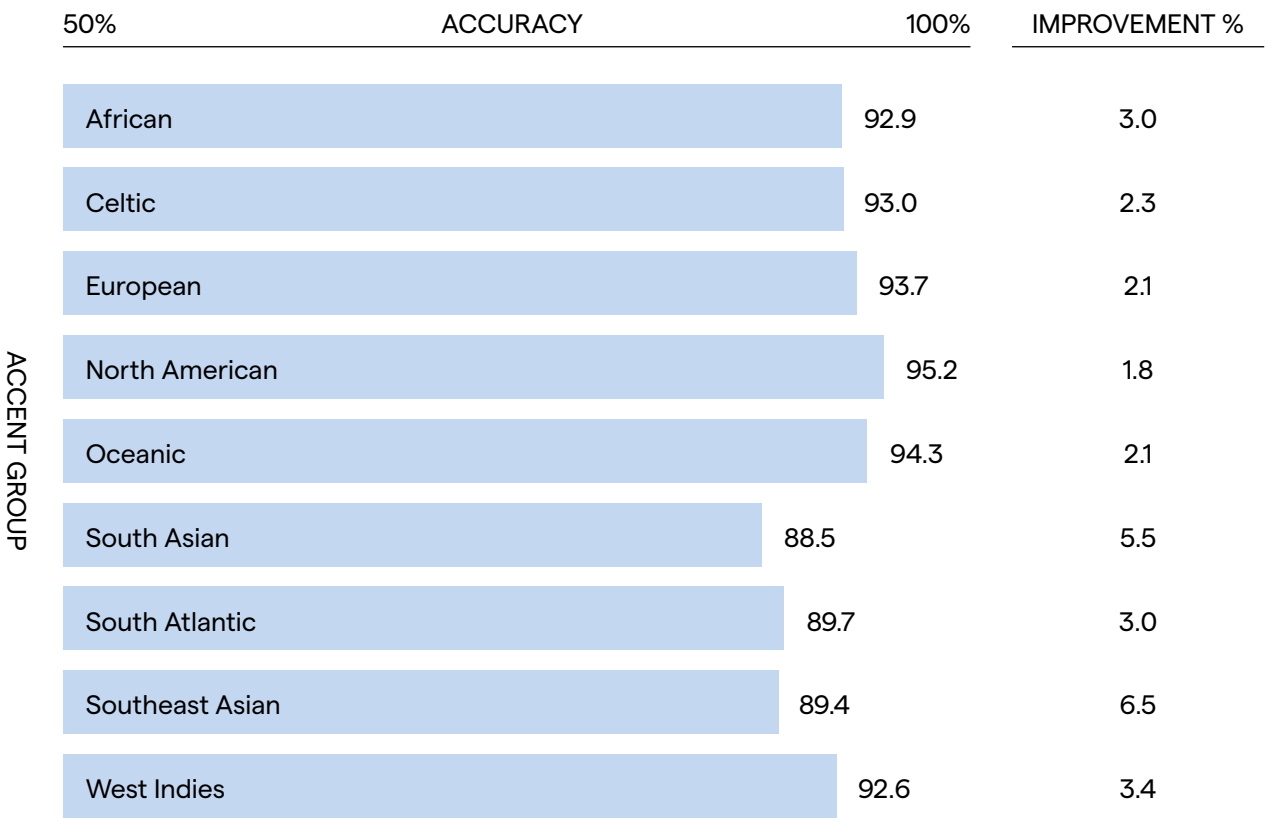
Try our latest technology now for free

Opening Doors. Understanding Every Voice.

The opportunities available now – through Autonomous Speech Recognition – start to look unlimited. With self-supervised learning at the heart of our approach, we can now expand to less commonly spoken languages and have greater confidence in audio with quality issues and background noise. Using accent-independent speech recognition, differing dialects in the same language will no longer be a barrier.

This approach also helps us tackle head-on one of the biggest issues in technology today: AI bias. We've seen positive results across the board when we've set our Autonomous Speech Recognition on different dialects, languages and age ranges.

Taking dialects as an example, you can see (opposite) how the new engine creates a massive improvement when tested against US English as a reference. Creating greater equity across different dialects, using English as the spoken language, is no longer an unattainable aspiration.



African	Southern African (South Africa, Zimbabwe, Namibia)
Celtic	Irish English, Scottish English, Welsh English
European	British English
North America	United States English, Canadian English
Oceanic	Australian English, New Zealand English
South Asian	India and South Asia (India, Pakistan, Sri Lanka)
South Atlantic	South Atlantic (Falkland Islands, Saint Helena)
Southeast Asian	Filipino, Hong Kong English, Malaysian English, Singaporean English
West Indies	West Indies and Bermuda (Bahamas, Bermuda, Jamaica, Trinidad)

*Accuracy improvement for speaker accent (US English as reference). Results are from Common Voice dataset (~7hrs of speech).

It's far from the end of the story – we believe there's always room for improvement – but for the first time, thanks to the leap forward of self-supervised learning and its seamless integration into our software, the fences are coming down.

To read more about how we're improving inclusion through self-supervised learning, read our free White Paper "Pioneering Greater Accuracy in Speech Recognition to Reduce AI Bias".

Keep in the Know

As we continue to publish our findings, we're inviting you to stay switched on to all the latest developments. Enter your email below and we'll send you an up-to-the-minute newsletter on all things "Speechmatics" - including more on our unsupervised network, how we're confronting AI Bias and where we are on our continuing mission to Understand Every Voice.

Find out more about our vision for
Autonomous Speech Recognition

There's plenty more to explore.

Stay in the loop with our latest news and updates

Information on the data used for testing in this whitepaper

Common Voice
Accent and age data are from Common Voice, an open-source project seeking to make a speech dataset freely available for anyone.