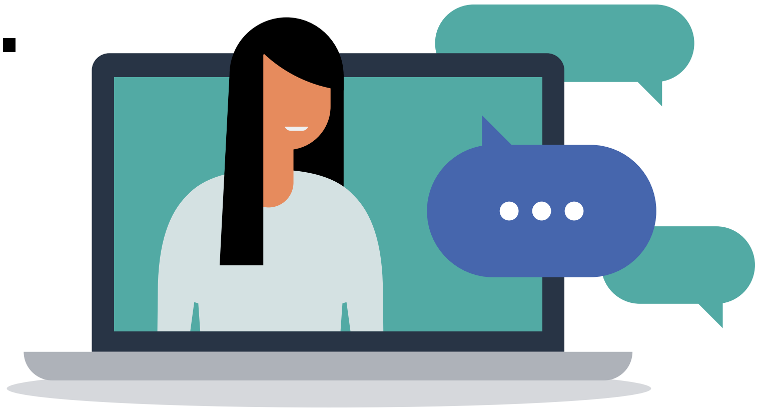
Speechmatics

# Opening Doors.

## Understanding Every Voice.



November 2021

WHITEPAPER

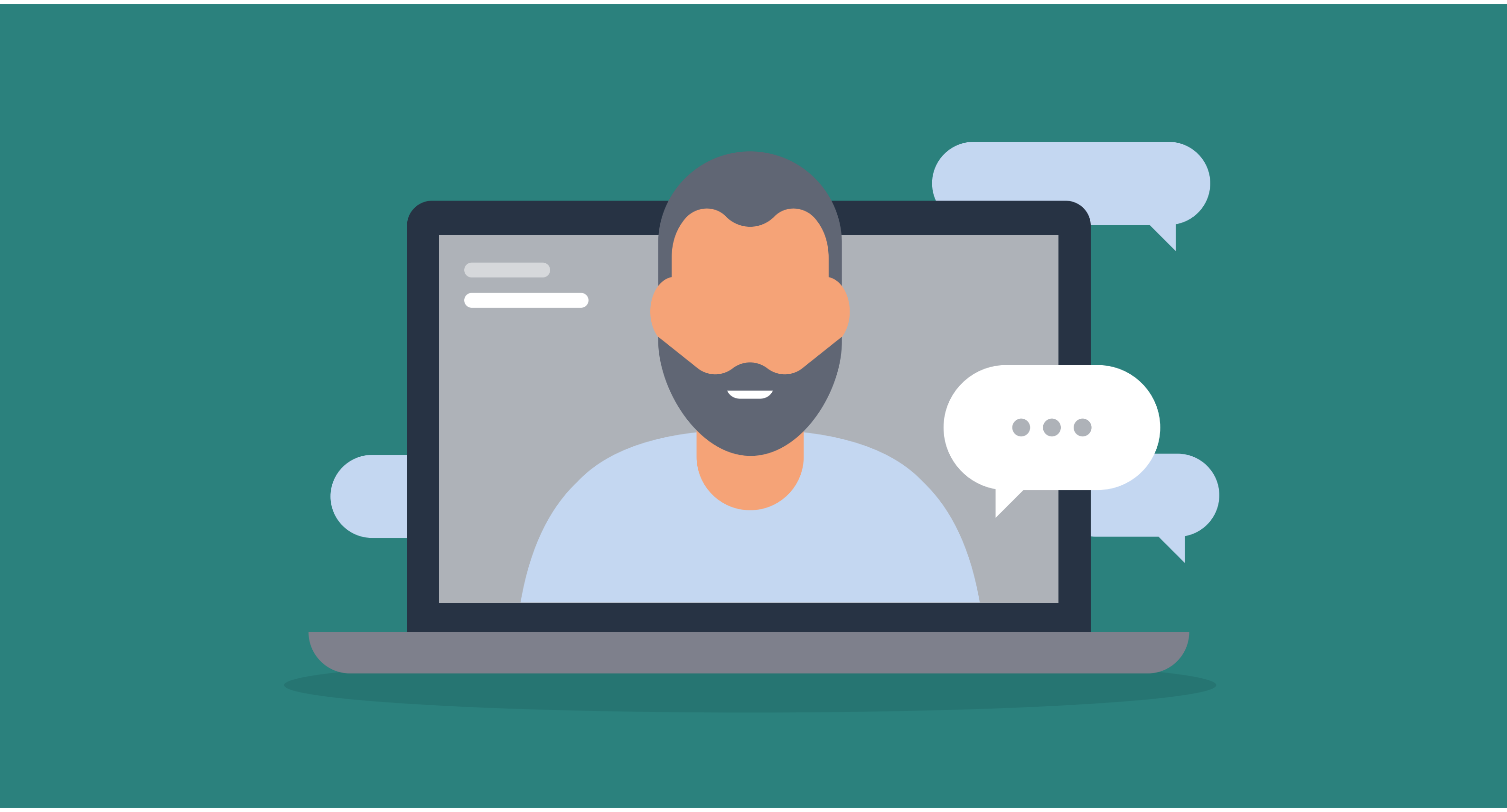
Speechmatics.com

# Pioneering Greater Accuracy in Speech Recognition to Reduce AI Bias

What gives your voice its individuality is as complex and varied as everything that makes you the person you are. It may be guided by your parents, your siblings, your friends and your education. The area you live in might play a huge role, as might your job and even the media you consume. But no **one** single contributing factor makes your voice yours.

Opinions are formed about us when we speak. We can be judged on how we sound and there are times when how we speak can either be misconstrued or completely ignored. When it comes to speech recognition, we believe this shouldn't be the case. When the technology is at its optimum, every voice should be treated the same.

But how does this play in reality? When it comes to speech recognition, is there an inherent bias in the system? Are some voices being understood and others bypassed? Are there voice cohorts being let down by technology that's designed to be non-judgemental?



At Speechmatics, we turn what someone has said into the written word for assistance, reference and analysis. If our results are inaccurate, it means someone isn't being heard and we're no closer in our mission to understand every voice.

To see where the problem lies, and more importantly to do something about it, we looked into different voices from a multitude of backgrounds, grouping them only to see if there were patterns where sectors of society were poorly served.

The results led to us taking immediate action, improving our accuracy across the board with a new way of thinking. A way that saw huge reductions in historical disparities, meaning those who were once left out are now, finally, given a voice in the room.

> **"It's critical to study and improve fairness in speech-to-text systems given the potential for disparate harm to individuals through downstream sectors ranging from healthcare to criminal justice.**
>
> **Allison Zhu Koenecke**
> Lead author of the Stanford study

## Admitting the Problem. Committing to a Solution.

In the world of Automatic Speech Recognition (ASR) there's been a danger that languages, dialects and ways of speaking that aren't the ones machines have been trained on (namely the default of white and well-educated) will be misunderstood.

In Spring of 2019, Stanford compared individuals whose ethnicity was identified as white, and African American voices, using large-name speech recognition software. The results were far from good. Using Google, IBM, Apple, Amazon and Microsoft, on average Stanford found accuracy levels of 81% when white people were talking. When Black people spoke, this plummeted to 65%. For Black men, it was closer to 60%.

It's impossible to argue a disparity so clear and well-articulated by data. ASR was failing Black speakers. But this gap can be found across other areas too, from dialects in every language through to age discrimination.

Here at Speechmatics, we accepted the problem as a truth and committed to addressing it. We know the data we train on isn't the only factor (biases can come from everywhere, with hiring and empowerment two major factors) but from a technological standpoint, data is everything.

We needed a new way of looking at how our engines are trained. We needed wider datasets. This forward-thinking led us from Automatic Speech Recognition onto our new journey towards Autonomous Speech Recognition.

> “
> **Speech recognition is a technology that will be used by each human on the planet every day. It needs to work for everyone. Period.**
>
> **Sam Ringer**
> Chief Machine Learning Engineer, Speechmatics

> **“ Our journey towards 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

## The Road to Autonomous Speech Recognition

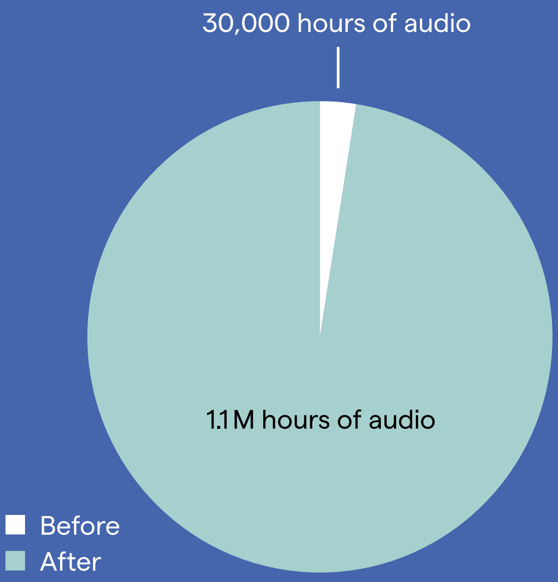
For too long commercial speech recognition has been at the mercy of the availability of labeled data. By bringing speech recognition to the commercial space, we felt it was imperative we didn’t also bring biases into language learning, call centers and the broadcast world.

When it comes to labeled data, some areas have been plentiful. In particular, clear, well-edited American English. Other languages, demographics, and lesser-quality recordings have been harder to come by. Put simply, machines have learned to recognize specific languages and particular accents in specific situations only.

Historically, manually labeling data takes time, effort and money. In speech-to-text, for every audio file with an annotated transcript there’s someone listening, pausing after a few words and manually inputting the words spoken. Therefore, the cost of transcribing hours of audio is vast.

Our solution has been to look at how we go about training from that data. To make a real step-change meant creating an engine that trains on both labeled and unlabeled data. For the very first time, using our new self-supervised models, the costly middle stage ceases to be a factor.

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. Every broadcast, every YouTube video, every podcast can be used for training, ready to grow and expand on new and varied voices.



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

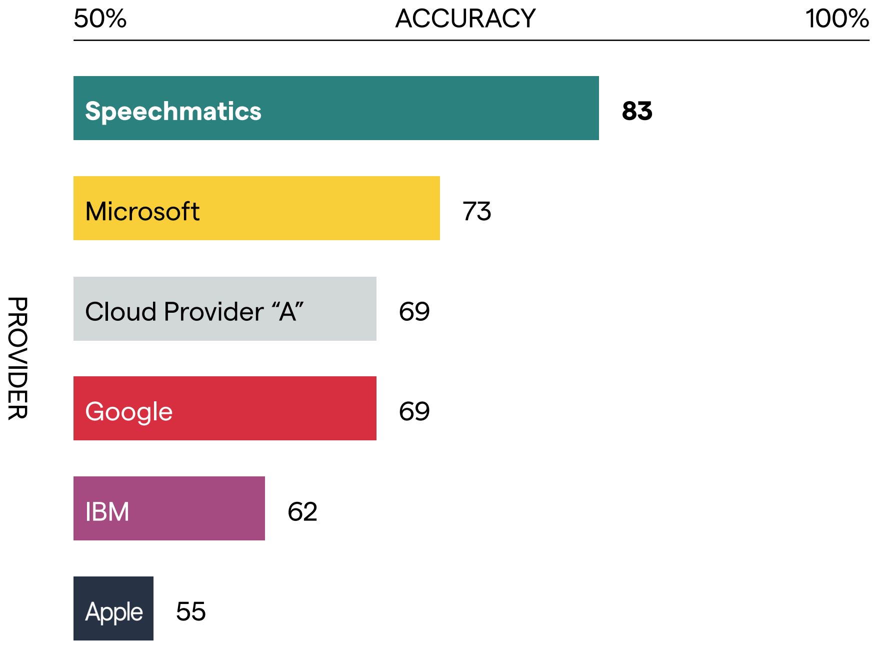
## Innovation and Accuracy

By leading the industry from Automatic Speech Recognition towards Autonomous Speech Recognition (ASR) and using the latest techniques in machine learning, including the introduction of self-supervised models, we can take accuracy to the next level.

This step-change means Speechmatics ASR learns by using larger datasets from a huge variety of sources that include both labeled and unlabeled data direct from the internet as well as specialist sources. And that means greater accuracy for all voices.

Using the Stanford study as a starting point (see below), we can see just how vast the disparity was. Speech recognition software performed so poorly for African American Vernacular English (AAVE), a couple of the major players barely made correct transcriptions for half the words spoken.

Yet, with the same recordings of Regional African American Language as the Stanford analysis, we found our Autonomous Speech Recognition (which uses self-supervised learning for training) was far and away the best at improving the accuracy amongst those who speak AAVE. It's still not close enough to say we've succeeded in our mission to understand every voice, but these initial tests are an incredibly encouraging step forward.

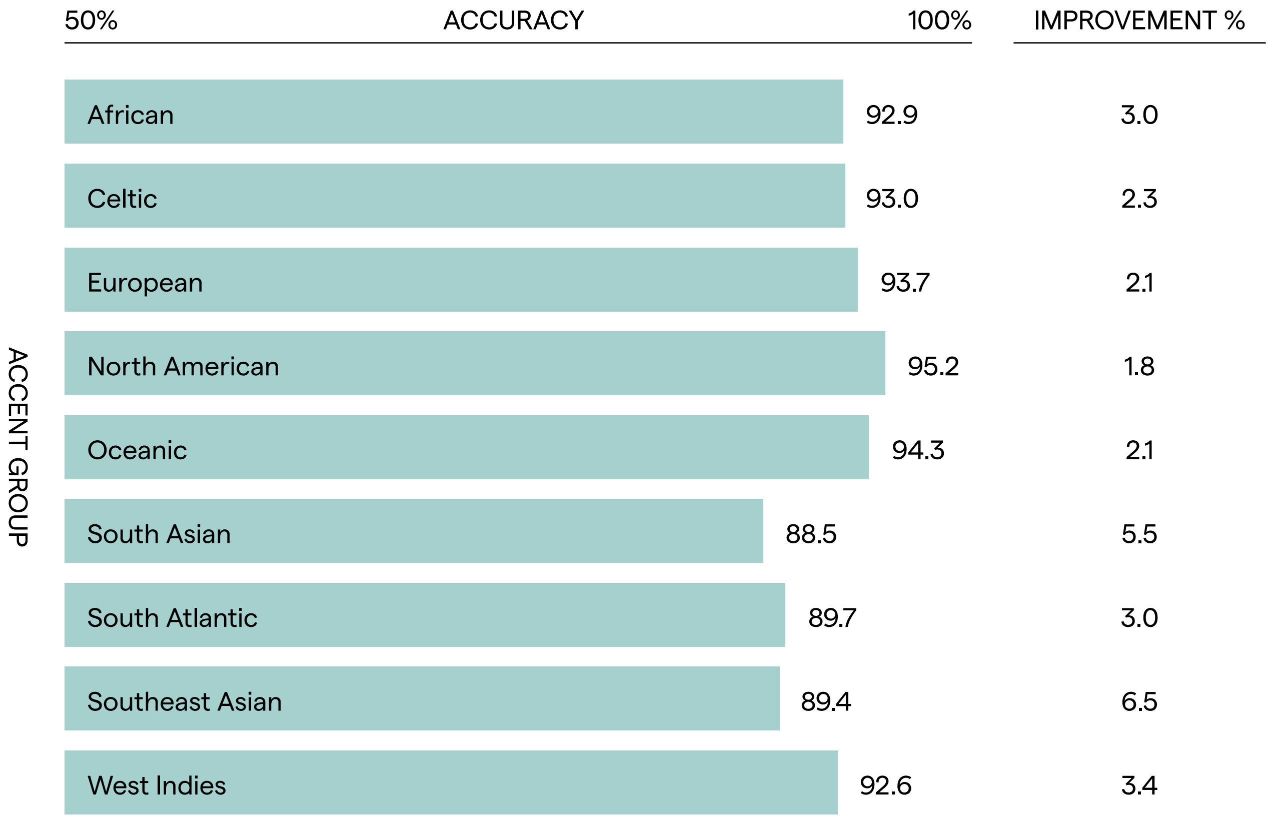


> **As a non-native English speaker, it is great to see this move toward Autonomous Speech Recognition and the accuracy it's providing for everyone in the world, no matter one's linguistic background.**
>
> **Benedetta Cevoli**
> Doctoral Researcher at Royal Holloway, University of London

## Accentuate the Positive

Traditionally, it isn't just AAVE speakers where the old ways of speech recognition have fallen short of parity. For all the reasons outlined above, speech-to-text software working from labeled data hasn't always performed well on different accents around the world.

Now with Autonomous Speech Recognition, we're seeing sizeable changes across a variety of different speech patterns, all using English as the spoken language.



| | |
|---|---|
| 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) |

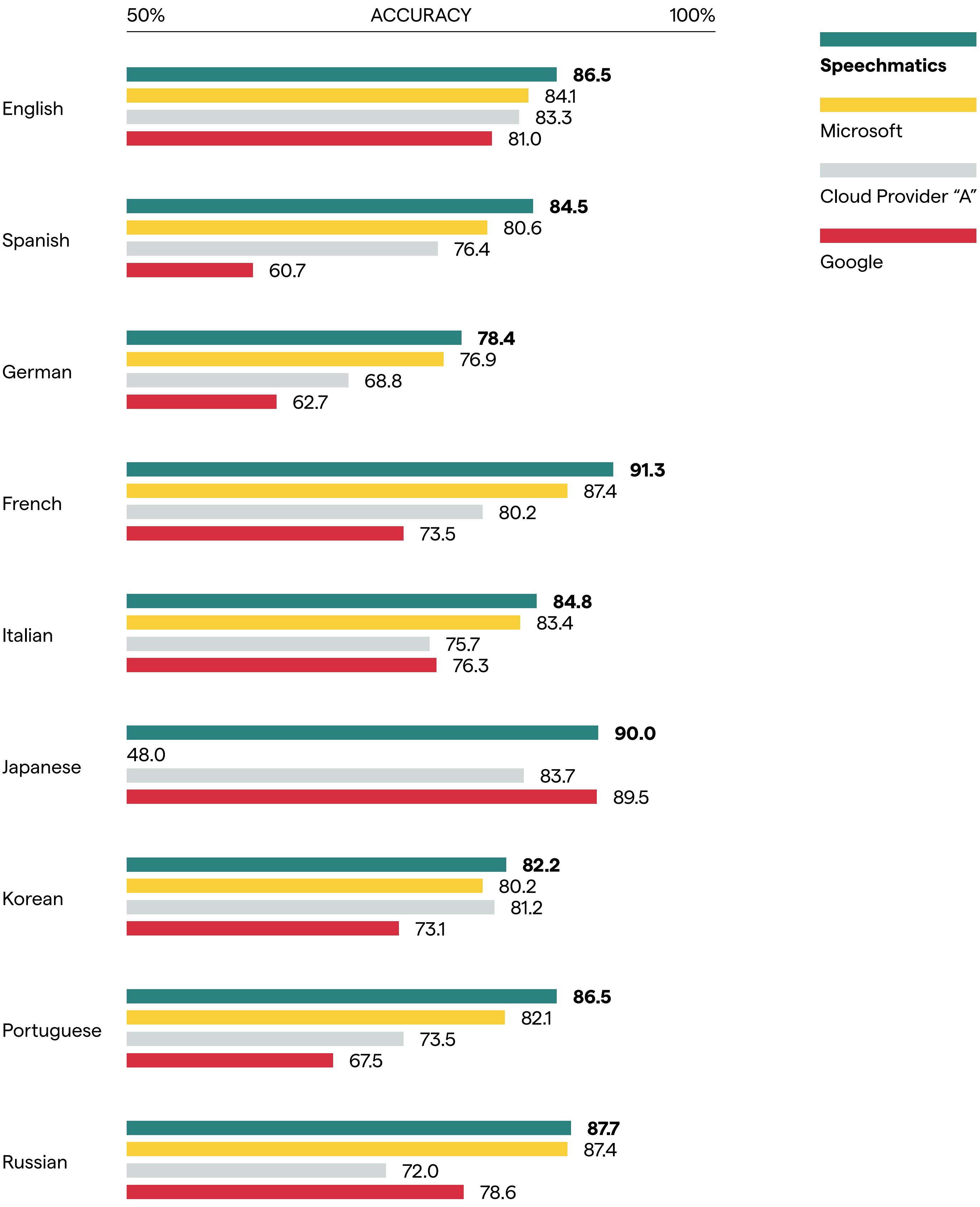
*Accuracy improvement for speaker accent (US English as reference). Results are from Common Voice dataset (~7hrs of speech).

# Global Goals

## Global Goals

So far, we've focused primarily on English speakers, but how does our move towards Autonomous Speech Recognition stack up when tested on languages around the world?

At the time of writing, we support over 30 different languages from Arabic to Turkish. While we won't stop until we cover every country, we're extremely pleased with our results on some of the most frequently spoken.

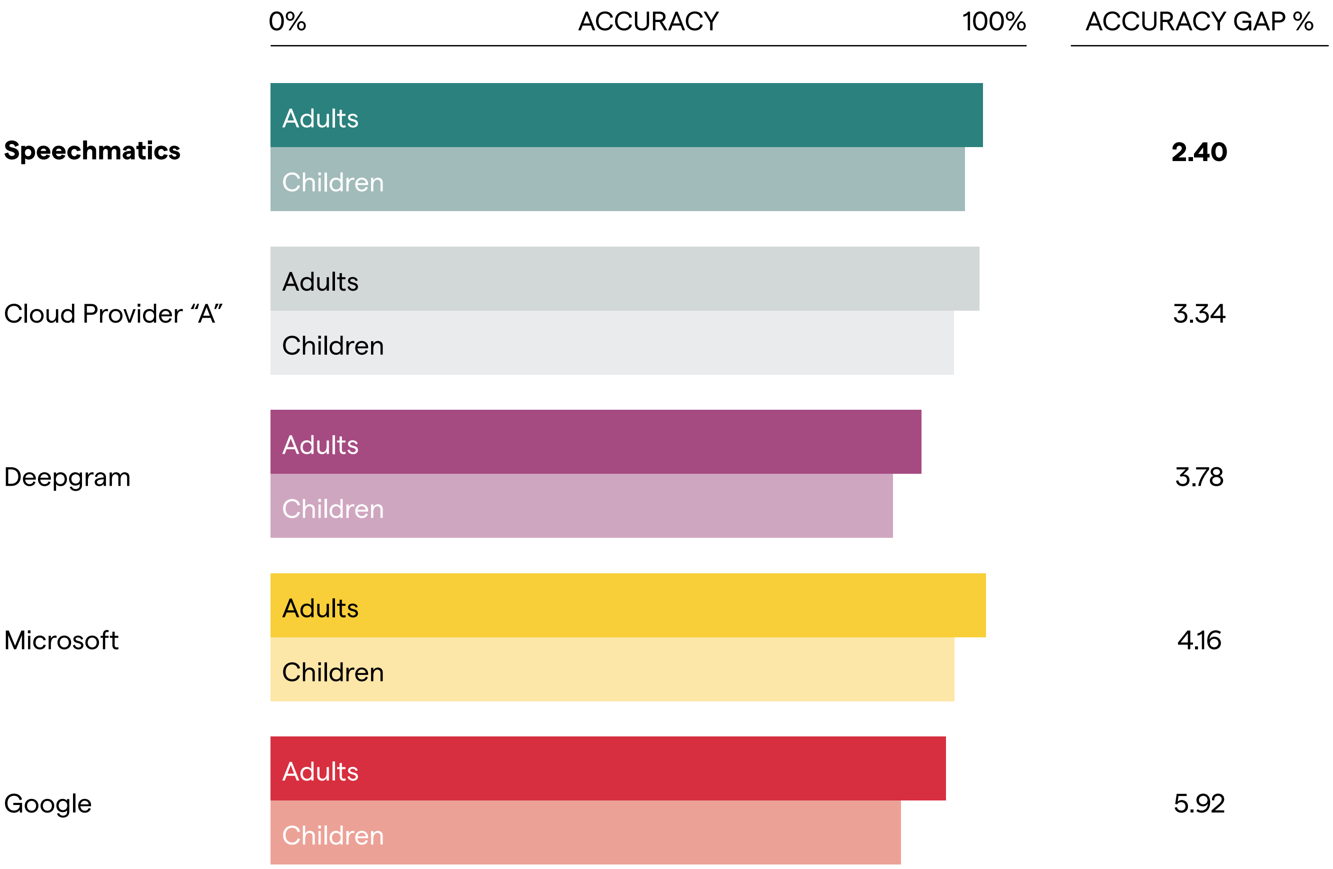


## Looking to the Future

The COVID-19 pandemic continues to change the world as we know it. It's still hard to predict what that change will look like in the years to come. However, one thing that became clear almost immediately was how essential eLearning is for children around the globe.

In the world of speech recognition – just as with accents – there's been a gap in accuracy between younger and older voices, with those under eighteen struggling to be understood as well as adults. Again, the scope of the labeled data to learn on has always been an issue, but now, working from labeled **and** unlabeled data, this bottleneck can be removed.

Our first round of results (using material from the open-source project Common Voice) has shown another fundamental shift in closing this gap.

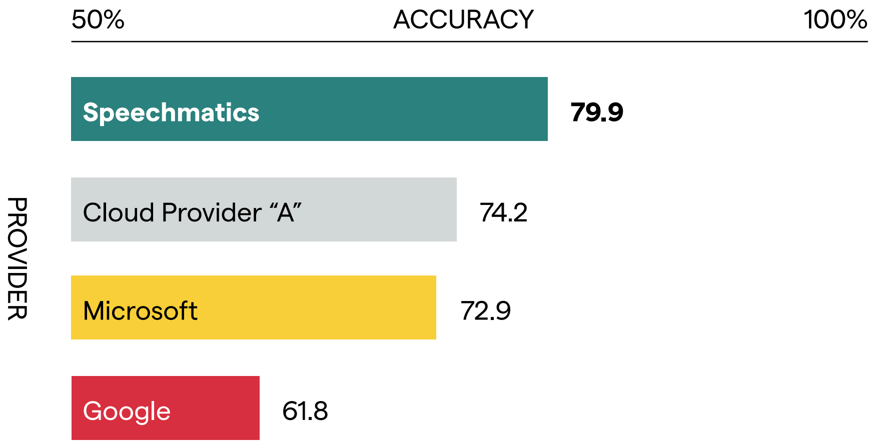


## Forward Moves Against Background Noise

An age-old challenge of transcription is the quality of the recordings we use. Material with background noise has always made it hard to turn speech into text. Even before ASR, poor quality recordings were often the biggest bugbear in human transcription.

To truly test our Autonomous Speech Recognition, we wanted to see how it would handle poor quality data. We manipulated 6 hours of audio taken from meetings, earning calls, online videos and a host of other real-world examples. We then added ambient sounds (phones, machine noise, other conversations, etc.) and tested our software against some of our competitors. To make it even more of a challenge we randomly played around with pitch, reverb and volume levels too.

As with age and accent, the results showed we were once again above our competition and making yet more promising steps in the right direction.



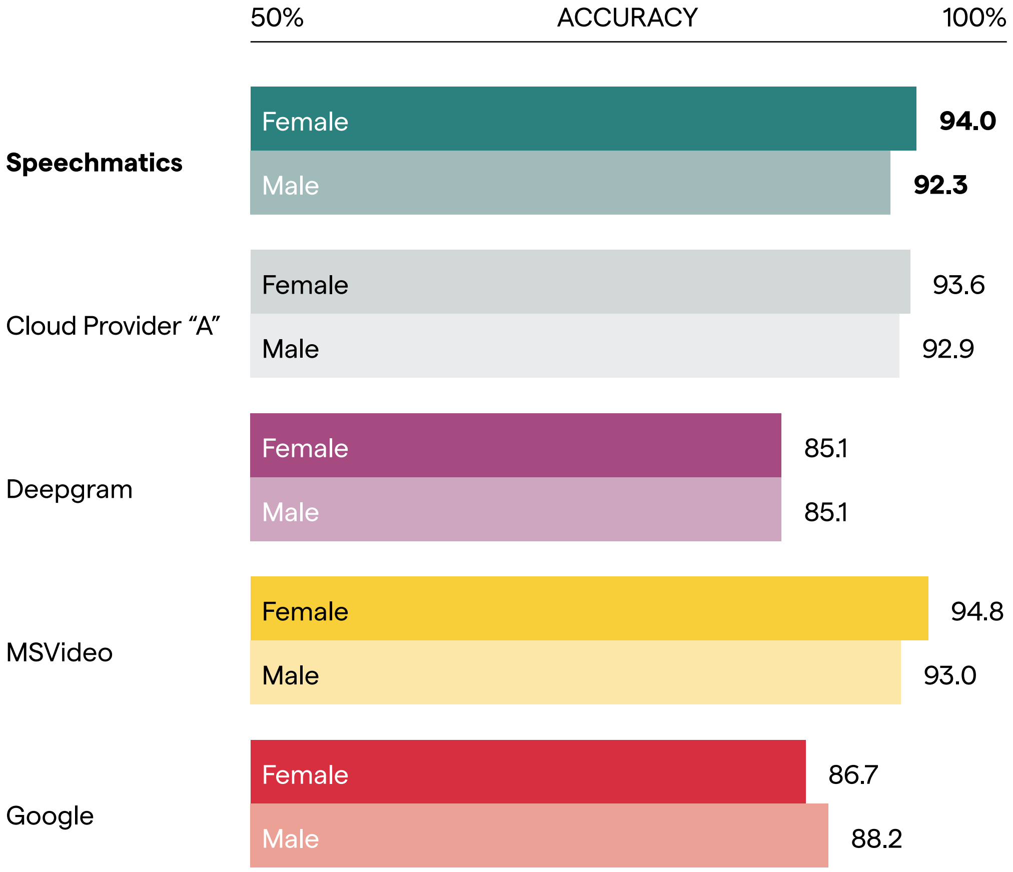
## A Question of Gender

Proving a disparity in the accuracy of speech recognition when it comes to gender has often proved difficult. Some tests have shown software delivers an accuracy gap that favours women, some tests show it favours men.

With so many other factors to consider – dialect, language, age, etc. – it might take time before we see notable trends. But by training with self-supervised models, we're hopeful any gaps will be minimal whichever way they fall.

As for the results themselves, while other factors have seen us ahead of our competitors by a distance, in gender we see a few of the other vendors in the same area for accuracy. If an organization trained on these opensource datasets, they'd achieve better results. This isn't something we do. All of the key results published in this paper are taken from tests where we've come to the data fresh.

We'll also continue to observe gender identity in ASR. As society and the world at large continues to evolve its thinking, we'll be doing everything we can to provide easy-to-use and accurate speech-to-text software for everyone – however we choose to identify.



## Challenges to Overcome

While self-supervised learning has allowed us to unlock hundreds of thousands of more hours of data to train on, because this is unlabeled data often we don't have the relevant information about the demographics of the voices we're using. Therefore, finding patterns isn't the easiest. Questions are also raised about who defines white and black and male and female.

In 2022, Speechmatics will be creating our own datasets, gathering as much information as possible where the voices themselves will define their identity.

Our mission has always been to Understand Every Voice. We're not there yet. Our bias still exists, but as the data shows, we've massively reduced it. The journey will be a long and winding one, but with these positive results the disparity between the past and the present, in terms of AI Bias, is the narrowest yet.

With self-supervised learning leading to Autonomous Speech Recognition, we've opened the door to an improved way of doing things. By doing so we've been able to bring the world the most powerful, inclusive and accurate speech recognition ever released.

We've more to do, but this bold, new journey has already begun.

# Keep in the Know

As we continue to publish our findings, we’re inviting you to stay switched on to all the latest developments. Sign up below and we’ll send you an up-to-the-minute newsletter on all things “Speechmatics” - including more on our self-supervised models, how we’re continuing to confront bias and where we are on our continuing mission to Understand Every Voice.

There’s plenty more to explore.

Speechmatics.com

Stay in the loop with our latest news and updates

**Research**

https://www.technologyreview.com/2021/05/20/1025135/ai-large-language-models-bigscience-project/

https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

https://www.scientificamerican.com/article/speech-recognition-tech-is-yet-another-example-of-bias/

https://www.pnas.org/content/117/14/7684

**Information on the data used for testing in this whitepaper**

(1) CORAAL
Corpus of Regional African American Language (CORAAL): ongoing collection of sociolinguistic interviews with individuals who speak regional varieties of African American Vernacular English (AAVE) conducted at different US sites (Washington, DC; Princeville, North Carolina; Rochester, NY).

VOC
Voices of California (VOC): ongoing compilation of interviews recorded across the state of California based on two main sites, Sacramento, and Humboldt County – only individuals whose ethnicity was identified as white were included in the analysis.

SBC
Santa Barbara Corpus of Spoken American English (SBC): large body of recordings of naturally occurring spoken interaction from all over the United States, representing a wide variety of people of different regional origins, ages, occupations, and social backgrounds – only individuals whose ethnicity was identified as white were included in the analysis.

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