Speechmatics

Our Vision for Autonomous Speech Recognition

Speechmatics.com

At Speechmatics, our goal has always been to understand every voice. To get us there, we know we need radical, forward-thinking change in the industry and Autonomous Speech Recognition is the answer.

When it comes to technological advances in speech recognition, the lack of availability of audio data has slowed down progress. Automatic Speech Recognition has always relied on consuming large quantities of data from a narrow set of speakers. This is typically achieved through the use of human-labeled data, such as audiobooks.

For years, speech recognition systems required human input on a vast scale. Be that scraping and processing data, building separate models for word pronunciations, or manually running various stages of neural network training. These processes took time and were often costly. Manually labeling huge amounts of training data has always been one of the major bottlenecks for improving Automatic Speech Recognition.

We say there's a better way. And that way is Autonomous Speech Recognition.

Fuelled by the introduction of self-supervised learning for our training, Autonomous Speech Recognition delivers a step-change in accuracy and inclusion by leveraging a wide range of voices using the scale and diversity of the internet. This approach takes us closer to understanding every voice regardless of accent, dialect, age, gender, or location.

Autonomous Speech Recognition is speech recognition produced with minimal human involvement. It's a vision and we're not fully there yet. But at Speechmatics, we aim high.

Making the Change For Those Who Need It

We know Automatic Speech Recognition requires human intervention and extensive labeled data, which is expensive, bias-creating, and error-prone. As innovators, we want to build systems and models that adapt to a much wider spectrum of voices to alleviate these issues. This in turn will lead to us providing much more accurate and inclusive transcription.

Customers of speech recognition technologies need and want much more than they had before. Their requirements have expanded beyond just the specific task of simple text transcription to include advanced capabilities such as punctuation, speaker diarization, global language packs, and new vocabulary with frequent builds.

Autonomous Speech Recognition enables us to coordinate the move from narrow speech-to-text to a more complete software solution for today's customers.

The Integral Components of Autonomous Speech Recognition

To elevate our Autonomous Speech Recognition, it requires four central parts; Self-Supervised Learning, Autonomous Data Collection, Autonomous Pipelines, and Accessibility Features.

Self-Supervised Learning

Sitting between supervised and unsupervised learning is the intermediate method of machine learning. Taking the best of both worlds, self-supervised learning typically uses artificial neural networks acquiring parts of unlabeled data to construct a supervised task.

There are a number of reasons we need to reduce our reliance on labeled data – from the cost and the time it consumes, all the way through to the way it can lead to bias. By training via self-supervised learning, our Autonomous Speech Recognition can bypass many of the usual pitfalls.

Autonomous Pipelines

Our tech stack can be pretty complicated. We need amazing pipelines to handle the complexity of training and deployment as we build Autonomous Speech Recognition for the world's diversity. To support our self-supervised technology and our autonomous data collection, we've rebuilt our stack from scratch in PyTorch and ONNX.

Autonomous Data Collection

Each year, as language constantly updates and evolves, new words enter our lexicon. A prime example, the vocabulary of the world changed in an instance in January 2020 when the word "COVID-19" first appeared.

Pronunciations, grammar, and idioms are always changing too, so we want our Autonomous Speech Recognition to benefit from Continuous Learning. This means learning from new data as it appears online and is ingested into our database, so our language can update automatically, without the need for human input.

Accessibility Features

For readability and accessibility, we require a number of extra parts to the software, including Neural Punctuation, Inverse Text Normalization, and Diarization. Without complicating things too much, the first adds punctuation, the second helps with formatting and the last helps identify who's speaking. Without these extra features, our output would be much more difficult to read and understand. With them, every customer's life gets a little bit easier.

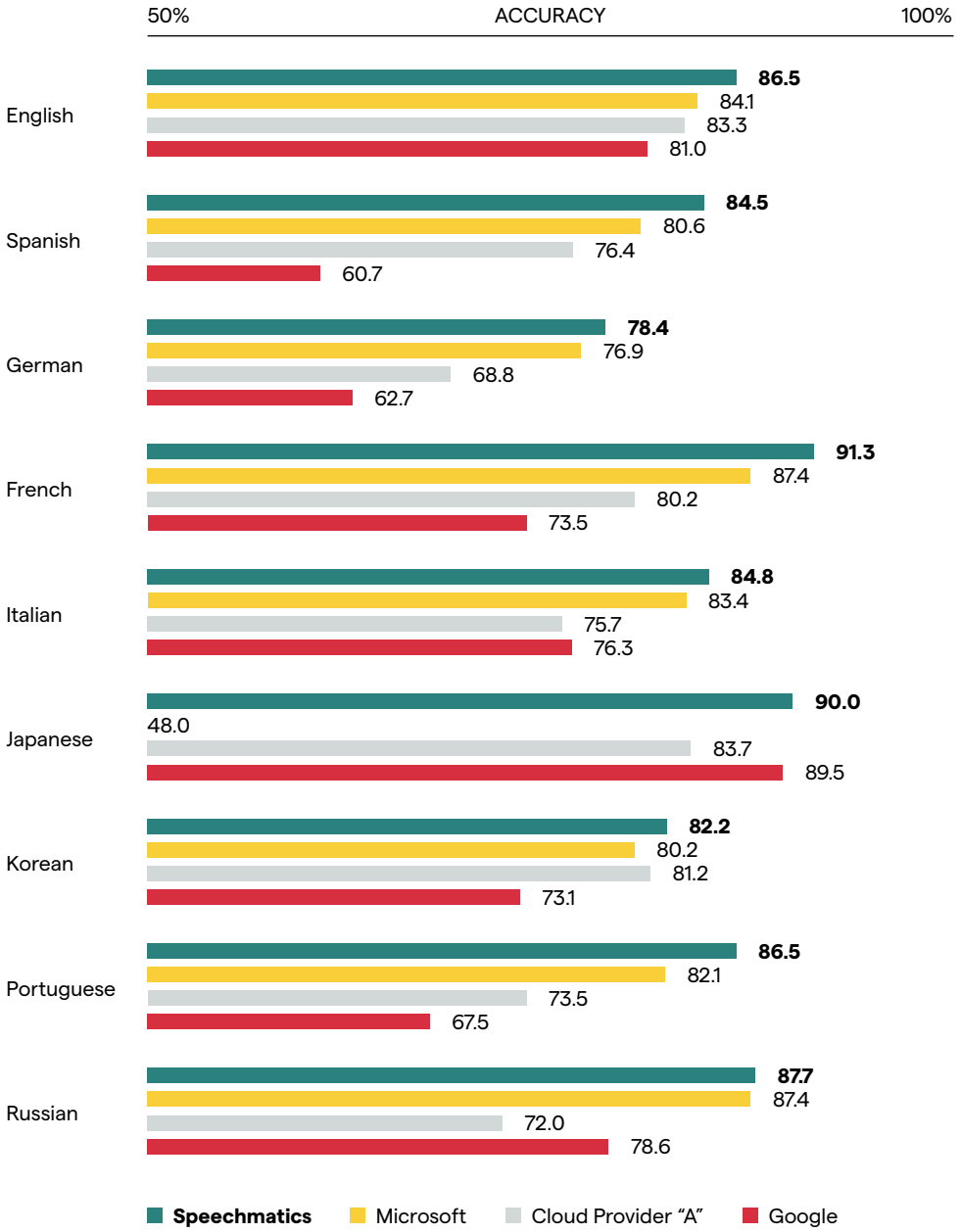
Arriving at an Autonomous, Accurate Future

We're proud of where we've arrived at with our Autonomous Speech Recognition, but we're not at the finish line yet. Our self-supervised learning is constantly improving and, due in large part to their complexity, our Autonomous Pipelines have been rebuilt from the start.

As for data collection, we're still relying on more human input than we'd like. But as we push the technology in new and exciting ways we've already seen evidence our accuracy has improved, outstripping our competitors and providing a more inclusive speech recognition experience for all.

As you can see from the data (right), when we compared our new Autonomous Speech Recognition to our competitors' speech recognition using a wide variety of different languages, we found ourselves head and shoulders above the rest. And we're confident, as the data we train on expands, these results will only continue to improve.

Speechmatics Autonomous Speech Recognition Vs Competitors Automatic Speech Recognition



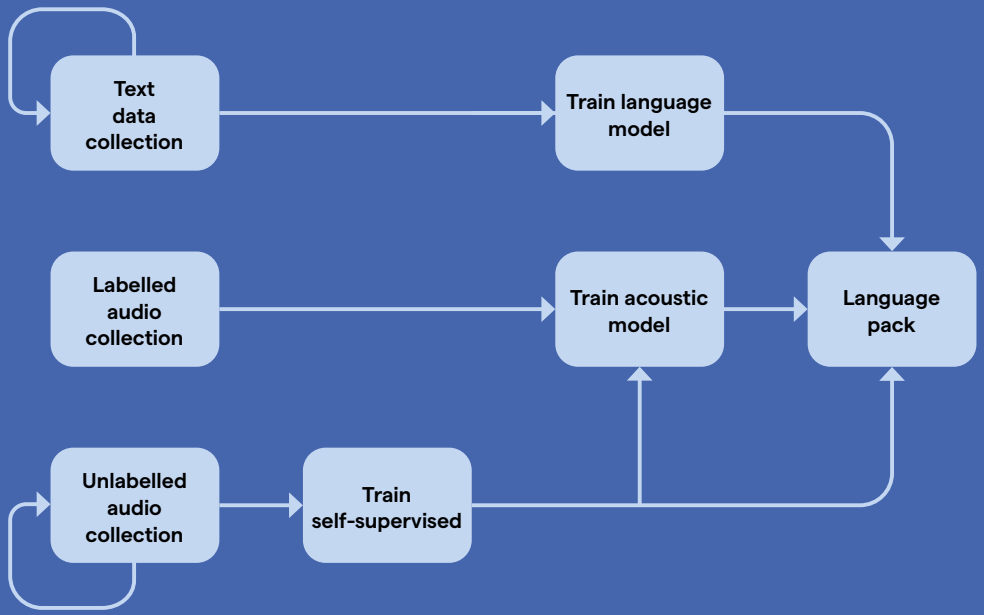
The Architecture Behind Autonomous Speech Recognition

Below, you can see some of the architecture behind Autonomous Speech Recognition. The high-level machine learning pipeline for Autonomous Speech Recognition includes three streams, each corresponding to the different models and types of data.

Text and unlabeled audio are continuously scraped from the internet to be used when training the Language Model and self-supervised model respectively. The continuous nature is represented by the looping arrows.

The acoustic model requires labeled audio data as it is trained in a supervised fashion so this data cannot be continuously collected. Once the three models are trained, they are packaged together in a language pack. The language packs are then used in the Autonomous Speech Recognition product that our customers receive.

The result is the most powerful, inclusive, and accurate speech recognition ever released.



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 move to Autonomous Speech Recognition, how we're continuing to confront AI bias and where we are on our continuing mission to Understand Every Voice.

There's plenty more to explore.

Stay in the loop with our latest news and updates