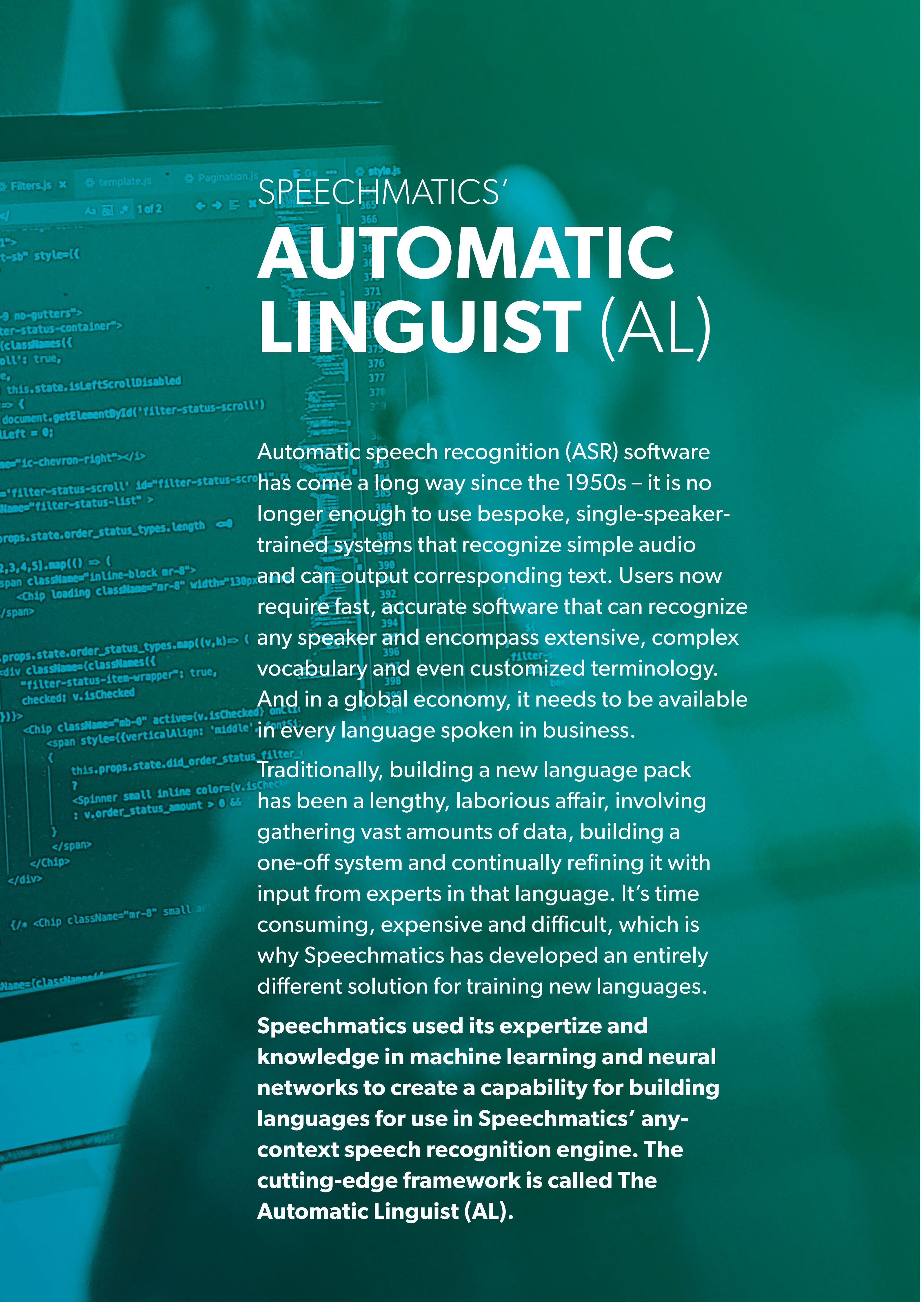
SPEECHMATICS

AL - THE AUTOMATIC LINGUIST

The breakthrough machine learning framework for training speech recognition language models

SEPTEMBER 2020

WWW.SPEECHMATICS.COM



SPEECHMATICS’

AUTOMATIC LINGUIST (AL)

Automatic speech recognition (ASR) software has come a long way since the 1950s – it is no longer enough to use bespoke, single-speaker-trained systems that recognize simple audio and can output corresponding text. Users now require fast, accurate software that can recognize any speaker and encompass extensive, complex vocabulary and even customized terminology. And in a global economy, it needs to be available in every language spoken in business.

Traditionally, building a new language pack has been a lengthy, laborious affair, involving gathering vast amounts of data, building a one-off system and continually refining it with input from experts in that language. It’s time consuming, expensive and difficult, which is why Speechmatics has developed an entirely different solution for training new languages.

Speechmatics used its expertize and knowledge in machine learning and neural networks to create a capability for building languages for use in Speechmatics’ any-context speech recognition engine. The cutting-edge framework is called The Automatic Linguist (AL).

BUILDING LANGUAGE PACKS

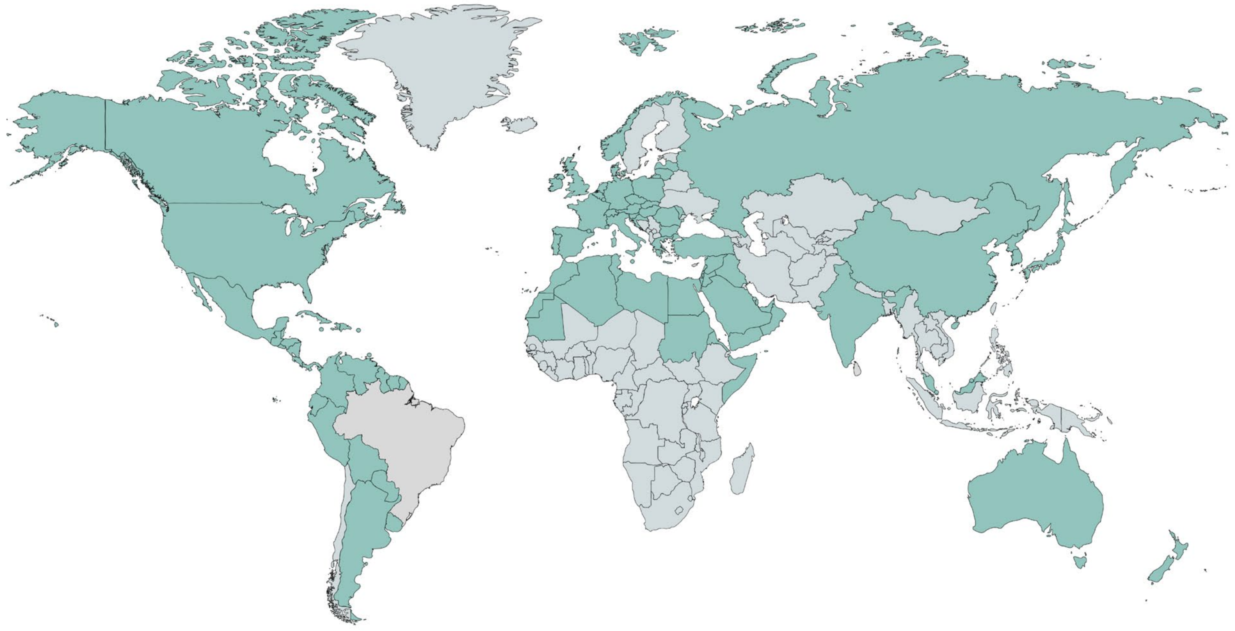
for the languages of the world

The world is increasingly connected and technologically dependent. Operating in a global marketplace requires technology to be accessible to users in any country, which means it must be usable in many different languages.

In many industries, speech is becoming the preferred way of interacting with technology, so speech recognition support must be available in a wide range of languages. Each of which would usually require a team to carry out extensive bespoke work purely to support that single language. We don’t believe this is the best way to build ASR in new languages.

In the early days of Speechmatics we focused on being the best in the world at English speech recognition. Then in early 2016 we shifted our focus to world languages, using our 30 years of R&D experience and unparalleled expertize in this rapidly evolving field to create a dynamic, internal framework for new language development.

Our languages portfolio



🟩 Core language coverage

Figure 1: Map of the world showing Speechmatics' language coverage.

Speechmatics' AL is our pioneering theory of how all the languages of the world work.

Whilst others look at the diversity of world languages, we look for patterns and order. We find patterns in letters that build up blocks of meaning; we find patterns in acoustics, as we all have the same physiology, and we find patterns in tones and other linguistic features. We find all this using machine learning, which means all the languages of the world fit into one framework for us.

This has the huge advantage that we only need to build our speech recognition code once, which we continually adapt for new linguistic features.

AL's automated framework enables us to build accurate language models for use in mission-critical applications.

Constantly evolving and improving through machine learning, AL then uses smart algorithms and cross-build learning to make each new language build easier and better quality. At the same time, our team of experts devise novel machine learning approaches when faced with unusual issues, rapidly assimilating new features and advancing them.

HOW DOES AL WORK?

The traditional way to build a language pack – life before AL

Before Speechmatics created AL, we built our language packs in a more traditional way. This involved:

Manual data handling

As low-quality data had a significant impact on the model quality, our speech recognition teams had to spend a great deal of time listening to, filtering and cleaning up data to make sure they were training only using good data.

This was a huge overhead, making new languages both slow and expensive to learn.

Specialized teams

Each of the many aspects of a language pack – from the acoustics of the language, to vocabulary and grammar – required a separate expert who focused solely on that aspect of the build.

Separating these aspects of the build into isolated teams meant that individual components did not always work together as well as expected.

It was also very difficult to balance resources, so bottlenecks would occur with no obvious way of relieving the pressure.

Linguistic expertize

Each new language we tackled needed a different linguistic expert to plan and manage the build as a new challenge, with specific measures put in place for that particular language.

These experts were only human. High levels of manual intervention could restrict the way the new language pack could be scaled up and generalized to other new languages.

The Speechmatics way – speech recognition reborn using cutting-edge machine learning techniques

As world experts in the traditional approach to building language packs, Speechmatics were looking for a more efficient solution. By adding value through improved speed and accuracy, and improving flexibility by using data to continually update all languages automatically, the AL approach – offering unlimited possibilities – was born.

A dynamic framework

AL was built to allow us to explore possibilities and then rapidly integrate the best techniques into the internal framework by using machine learning.

By using a standard framework for the development of all our language packs we can easily compare what differences new techniques make and establish which ones will keep us moving forward. This delivers our partners and customers highly accurate languages for use in their mission-critical applications.

Generic solutions to linguistic problems

We do not rely on linguistic expertize for every language. Instead, when we come across a linguistic problem we devise novel machine learning approaches to make the solution as generic as possible, so when we come across a similar problem in another language we don't need to solve it all over again.

Single pipeline

AL was built to develop and improve the Speechmatics language models, dealing with all aspects of a language pack. This means all the pieces of the build are created together and there are no surprises at the end when it all comes together.

This also means we have a holistic approach to building a language pack and can assign resources flexibly to meet the requirements of each new build.

Intelligent automation

While automating, we developed groundbreaking machine learning algorithms to distil our human expertize, and we will continue to improve and adapt this solution moving forward.

Cross-build learning

Starting a new language pack can feel like a mountain to climb. So we have developed techniques using machine learning and neural networks that mean we can always start from a well-established basecamp. This means learning across builds in different languages. AL learns from its previous builds, while the Speechmatics experts extract, streamline and improve on any new features to make future builds easier, faster and more accurate.

Accuracy of speech recognition

Speech recognition is a powerful tool, but only if it is accurate enough to fulfil your needs.

Expectations and operational requirements have become increasingly demanding over time, and our solution continues to provide reliable results and improvements. We are proud of the accuracy of our speech recognition, so we undertake regular comparisons against other providers to make sure we consistently lead the pack.

We do this by creating test sets of audio files and 'ground truth' matched transcripts of these audio files created by human experts. The test sets comprise 4 hours of audio files that fit customer use cases – a combination of broadcast data and contact center recordings.

We then run these test sets through our speech recognition systems and those of other providers, and calculate the accuracy of the output (see figure 2: 'One measure of accuracy').

As you can see in the comparison graphs (see figure 3), we outperform other providers in a wide array of languages. In many of these cases we do not have native speakers working on these languages – we instead rely upon AL's algorithms and cross-build learning.

We are careful to match these test sets to realistic use cases and there are many ways to measure accuracy. Many (in particular, academic) accuracy comparisons use very limited test data that give unrealistically high accuracy rates. This is typically because the data is much cleaner (has less noise) than is generated in real life, and the systems used in these tests are highly tuned to be specifically good on those test sets.

Our systems are instead trained and designed to work on a wide range of real world applications and our test sets reflect this too, making these comparisons fair, honest and as relevant as possible to your use cases.

Figure 2:

One measure of accuracy

In speech recognition the standard method of computing accuracy is to compare a 'ground truth' reference text to the output of a system. First count the number of times the system made a mistake (inserted, deleted or substituted a word); we can call this E. Then count the number of words in the reference text; we can call this N.

The accuracy as a percentage is then:

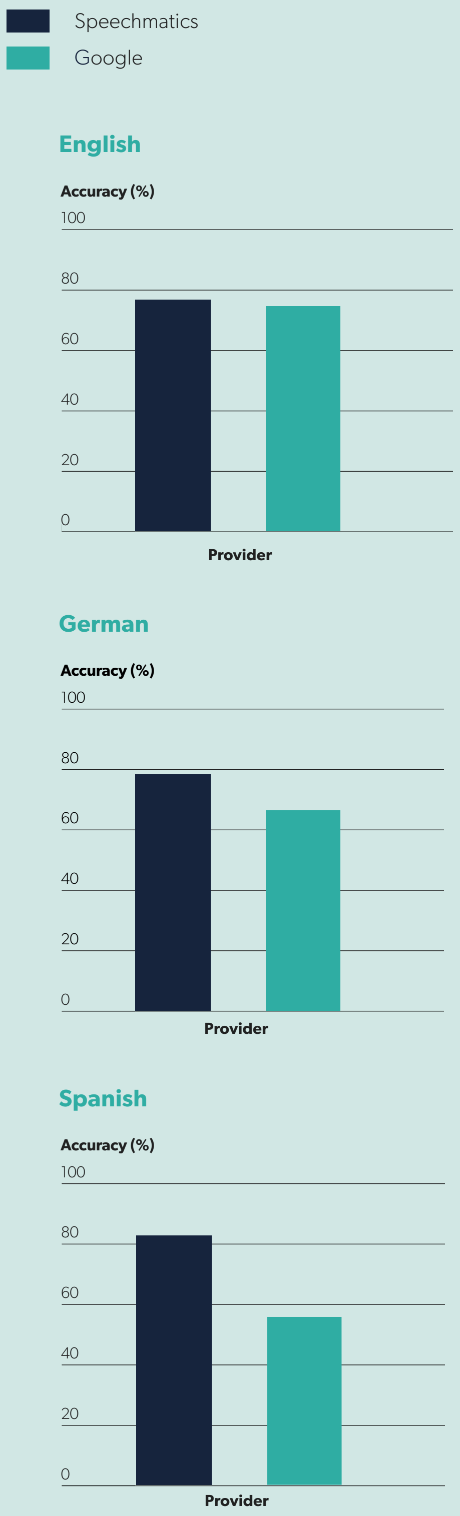
$$\text{Accuracy (\%)} = \frac{100 * (N - E)}{N}$$

For example, if your reference text said, 'The cat sat down' and your ASR output was 'The dog sat down' you have made one mistake, so E=1. There are 4 words in the reference so N=4. The accuracy is therefore:

$$\frac{100 * (4 - 1)}{4} = 75\%$$

Figure 3:

Graphs showing the percentage accuracy of Speechmatics against Google.



Figures taken February 2020. Test sets comprise of content from a variety of domains such as news, entertainment, and compliance.

Case study:

Red Bee Media

Red Bee Media, Access Services, specializes in delivering accessible content to make sure that people with hearing or sight impairments are able to get the most out of television content. Whilst regulated live content will continue to require skilled subtitlers, subtitling of pre-recorded content can be made more efficient using automatic speech recognition (ASR) technology. Partnering with Speechmatics, Red Bee Media is taking the most accurate ASR on the market and channelling it through its bespoke production workflows to provide same-language subtitling for pre-recorded content.

"At Red Bee Media, we systematically assess ASR solutions in order to know we are using the best ASR for the job. We've been really thrilled to have seen an average of 22% improvement in accuracy using Speechmatics' latest models across all our core languages but primarily in Global English and Spanish, on top of the already excellent base. These improvements will allow us to produce more subtitles for less and we are fast-tracking the new models in to production as fast as we can."

– Hewson Maxwell, Head of Technology, Access Services at Red Bee Media.

Speed of build

AL enables us to build and improve languages more efficiently than ever before.

This means we can provide our partners and customers with more efficient updates and accuracy improvements than the traditional approach previously allowed.

New languages

Our aim for AL was to be as language independent as possible, enabling it to generalize across multiple builds and languages.

However, when we encounter a language for the first time there is a chance it will have some attributes that mean we must update our techniques before we can start a build. We will focus on finding a general solution to the problem so we will not have to redo this work for similar languages in the future.

We focus on providing regular updates and improvements to our existing languages so our customers always have the best speech technology available. However, thanks to our framework AL, new language builds often lead to improvements of our existing language models too.

The exact amount of time it takes to build a new language pack depends on several factors, including:

Data acquisition

We use data from a mixture of commercial speech corpora, internally built corpora and/or customer supplied data. For different languages these may take different lengths of time to obtain and prepare for use in AL.

Data quantity

In simple terms, more data will take longer to build on as there are more numbers to crunch (though more data does normally give better results – see page 11 for further information).

Hardware availability

AL is based on complex cutting-edge algorithms and hence needs a lot of computational hardware to work with. If we build numerous different language packs at once, each one will take that bit longer as it will share the hardware we have available.

Build time is very dependent on the variables mentioned and is best demonstrated through examples. Most builds are somewhere between the two examples given in the case studies below.

Case study:

Speed of building Hindi

We were challenged to build Hindi from scratch with no native speaker knowledge within the company.

1. We already had some data available, which saved time
2. It was a small amount of data, so that also meant the build was fast
3. This was a speed challenge, so we cleared other builds from our hardware
4. We made some minimal changes to accommodate Hindi, which took us around two days

In the end we had a working system within **one week** of starting.

Case study:

Project Omniglot

The Speechmatics team put AL to the test to see how many languages the framework could learn in just 6 weeks.

The innovation challenge stress-tested AL to prove how well the machine learning behind AL worked. AL built 46 working research languages in just 6 weeks.

The smart algorithms behind AL enabled the team to experiment with building languages that would otherwise have been uneconomical to build.

Whilst we have 15 core languages that are more efficiently improved and applicable to most use cases and customer applications – we also offer 59 research languages thanks to Project Omniglot.

A word about data

Data is important for the success of any machine learning project and AL is no different. However, our training algorithms allow us to use much less data and, thanks to our filtering methods, AL can build a language pack on noisy data.

1. Quantity

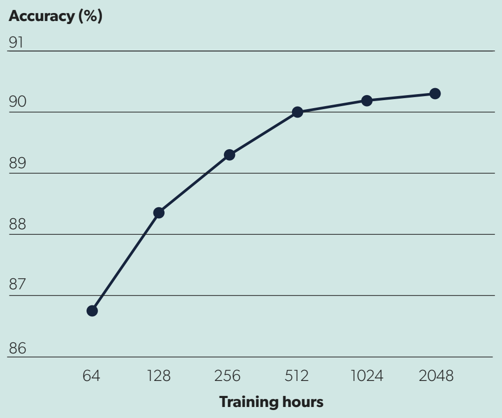
Traditionally, more data makes for better quality language packs. Every speech sample, even from the same speaker, is slightly different, and the more variations the system has been trained to expect, the better.

As you can see in figure 4, more data provides greater accuracy. However, there are diminishing returns and eventually a plateau is reached past which there are no further significant improvements.

It is possible to beat this plateau using more, quality data.

Figure 4:

Graph showing the relationship between number of hours of training data a language pack was built on and recognition accuracy, for one data set.



2. Quality

In machine learning there is a common understanding that you get out what you put in – good data leads to a good model.

Quality of data can mean many things in the context of speech recognition, for example, levels of background noise in audio, accuracy of transcripts to train on, and consistent use of spelling and grammar.

However, high-quality data is not always available, and we must often make do. This is where, in the traditional world of language modeling, manual data cleaning would take place.

We replace this with automated methods to filter and clean our data. This both improves results and reduces build times (by allowing us to train on less data) as well as reducing development time required.

Measuring quality is different to measuring accuracy. An example of this is shown in figure 5. Here we plot perplexity (a measure of quality of language models, where a lower value implies a better quality model) against the number of words we have trained our language model on. In one case we added more and more words randomly from a large (8 billion words) corpus; in another case we first filtered the large corpus and only added the highest quality data. As you can see, the filter allows us to 'beat the curve' despite using less data.

We use similar techniques throughout AL, replacing the traditional manual data handling methods to improve performance and increase speed of builds.

3. Domain specificity

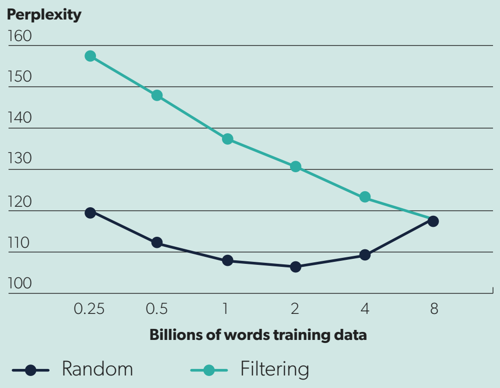
As we discussed previously, certain data has idiosyncrasies related to vocabulary or acoustic environments. In these cases well-matched training data will give improved results.

Because of this, we seek out data that is well matched to known use cases and tune our filters to include it.

As more use cases for our products are revealed, we will continue to refine our data parameters to match how we know our users will consume our language packs.

Figure 5:

Graph showing how perplexity (a measure showing language model quality – lower is better) improves with more data, and improves even more quickly if you filter the data to only include the best quality.



THE CHALLENGES AHEAD

AL has come a long way, but we have even bigger plans for the future.

What's next for AL?

Using AL, we have built an English accent-agnostic language pack called Global English. Regardless of accents, Global English can recognize and transcribe any audio – especially long-form audio – featuring English speakers. See our whitepaper "The Speechmatics approach to Global English" for more information.

We are at present looking to teach AL 'difficult' languages so that it can learn common solutions which will in turn improve existing language packs.

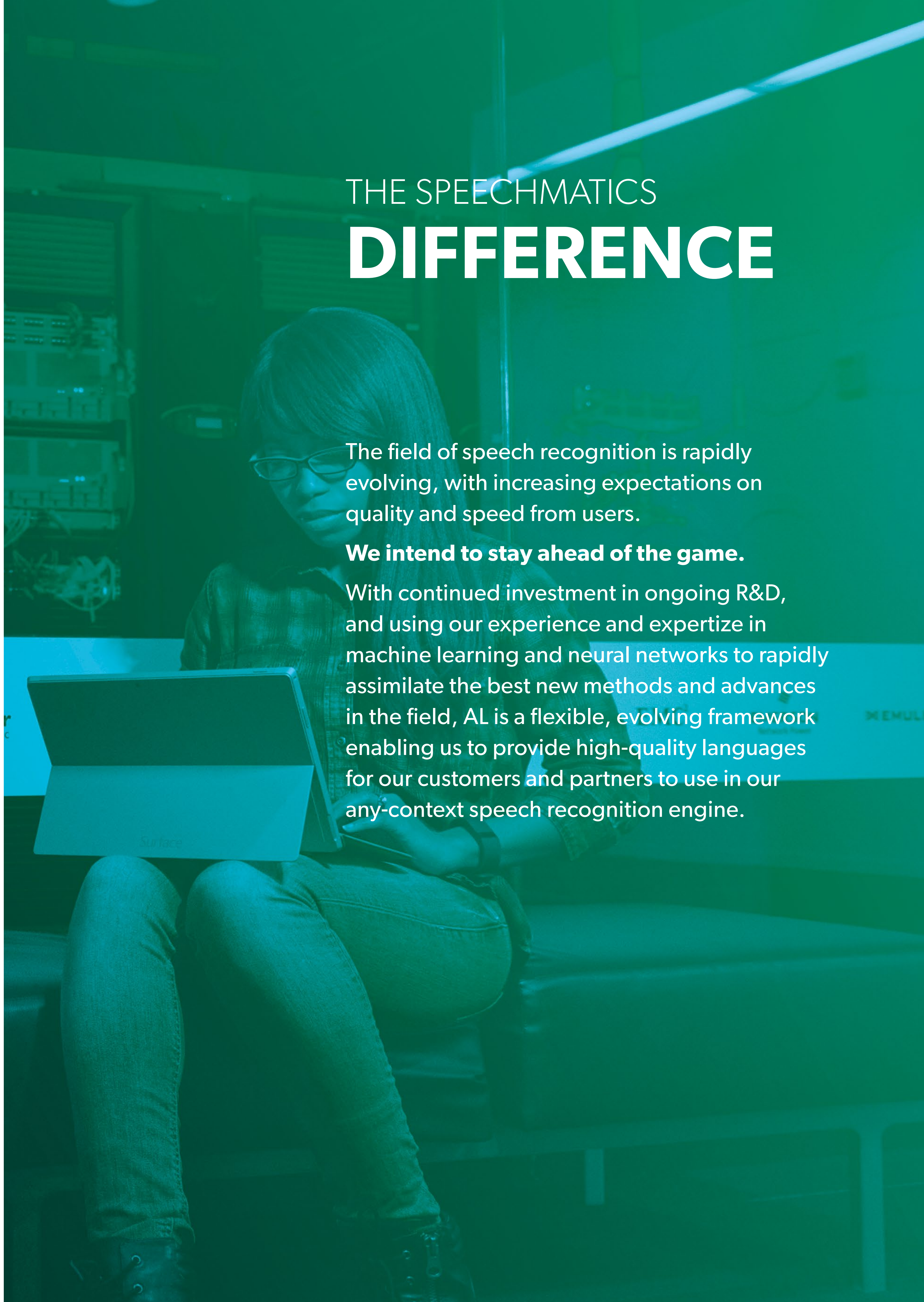
User needs are always increasing. With globalization and the ready availability of technology, more and more people want to be able to communicate, and speech is the most natural way to do that. This means a rapidly expanding market for speech recognition in a growing number of languages.

Taking speech recognition everywhere

The number of use cases for speech recognition is ever increasing as people recognize its potential.

We aim to support as many of these use cases as possible. This means continuing to update AL to produce language packs that work in a broad range of domains.

It also means making sure AL can produce language packs that can be consumed in various different ways. The vast number of language packs that AL produces appeals greatly to our existing customers and has been crucial to winning new customers across multiple platforms and for many use cases.



THE SPEECHMATICS DIFFERENCE

The field of speech recognition is rapidly evolving, with increasing expectations on quality and speed from users.

We intend to stay ahead of the game.

With continued investment in ongoing R&D, and using our experience and expertize in machine learning and neural networks to rapidly assimilate the best new methods and advances in the field, AL is a flexible, evolving framework enabling us to provide high-quality languages for our customers and partners to use in our any-context speech recognition engine.

SPEECHMATICS

296 Science Park
Milton Road, Cambridge
CB4 0WD United Kingdom

www.speechmatics.com