SPEECHMATICS

AL – THE AUTOMATIC LINGUIST

The breakthrough machine learning platform for global speech recognition

SEPTEMBER 2017

Introducing 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 recognise simple audio and can output corresponding text. Users now require fast, accurate software that can recognise any speaker and encompass extensive, complex vocabulary and even customised terminology. And in a global economy, it needs to be available in every language of the world.

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. Traditionally, it's time consuming, expensive and difficult, which is why Speechmatics has developed an entirely different solution.

Using cutting-edge ASR techniques, Speechmatics' AL uses machine learning and neural networks to learn any language in the world in as little as a week.

BUILDING LANGUAGE PACKS

in the languages of the world

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

Speech is becoming the preferred mode of input in many domains, 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 expertise in this rapidly evolving field to create a dynamic framework for new language development.

What is a language pack?

A language pack is the term we use to describe the collection of models that are required to perform speech recognition with one of our systems.

Each different language requires a different pack; we also sometimes have different packs for different accents/use-cases.

Our languages portfolio is constantly expanding.

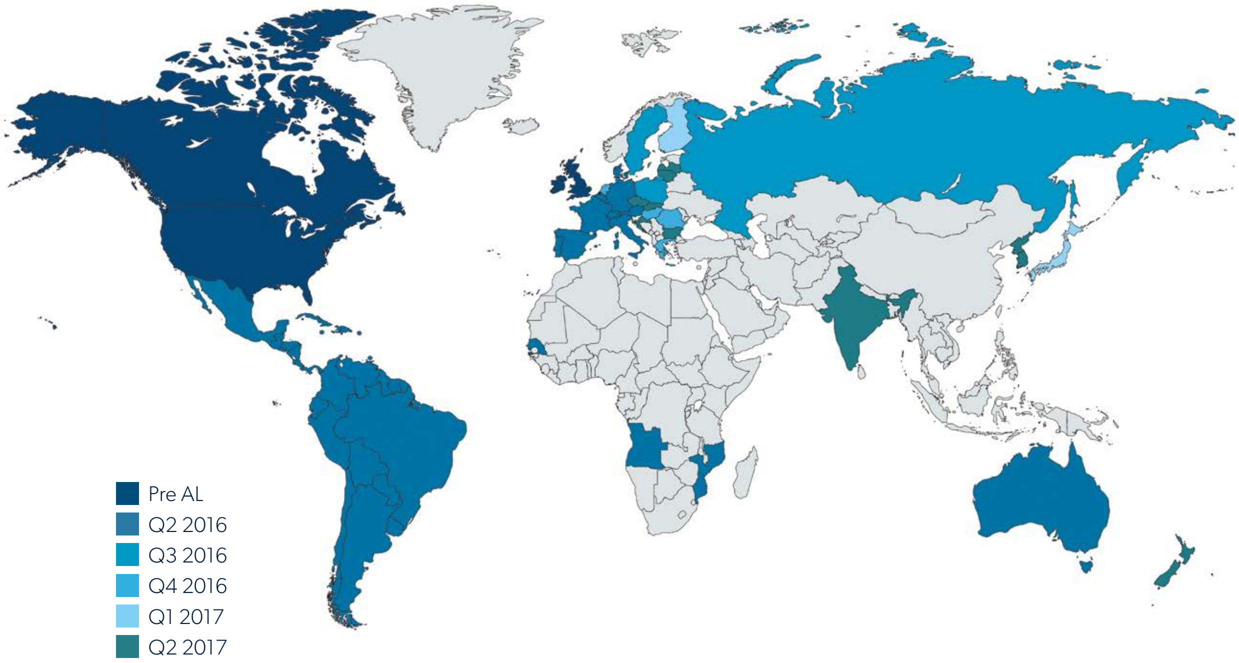


Figure 1: Map of the world with colour coding corresponding to when we first developed support for the language in that country in our first 15 months of building languages.

Speechmatics' AL is our pioneering new 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 even find patterns in tones and other linguistic features. We find all of this using machine learning, which means that 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, and we only have to update it for new linguistic features. What's more, we find cross-lingual patterns, so we can work with minimal data from a new language (sometimes just tens of hours) in order to build an effective speech recognition system.

AL's automated framework enables us to provide fast, accurate speech recognition in many languages easily, quickly and reliably.

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, ensuring our core ASR increases in accuracy across all languages. 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 to ensure that AL stays ahead of the curve.

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 the good data.

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

Specialised 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 separate 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 could be formed by more difficult aspects of the build, with no obvious way to relieve the pressure.

Linguistic expertise

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, and high levels of manual intervention could restrict the way the new language pack could be scaled up and generalised 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 is built to allow us to explore possibilities and then rapidly integrate the best techniques into our way of doing things by using machine learning.

By using a standard framework for all our language packs we can easily compare what difference new techniques make and establish which ones will keep us moving forward.

Generic solutions to linguistic problems

We do not rely on linguistic expertise 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 has been developed as a single project, 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 is all put 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

Automating the actions we were performing as part of building language packs, we developed ground breaking machine learning algorithms to distil our human expertise, 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 base camp. 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 fits customer use-cases – a combination of broadcast data and call centre 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: 'How do you measure 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. 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:

How do you measure 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:

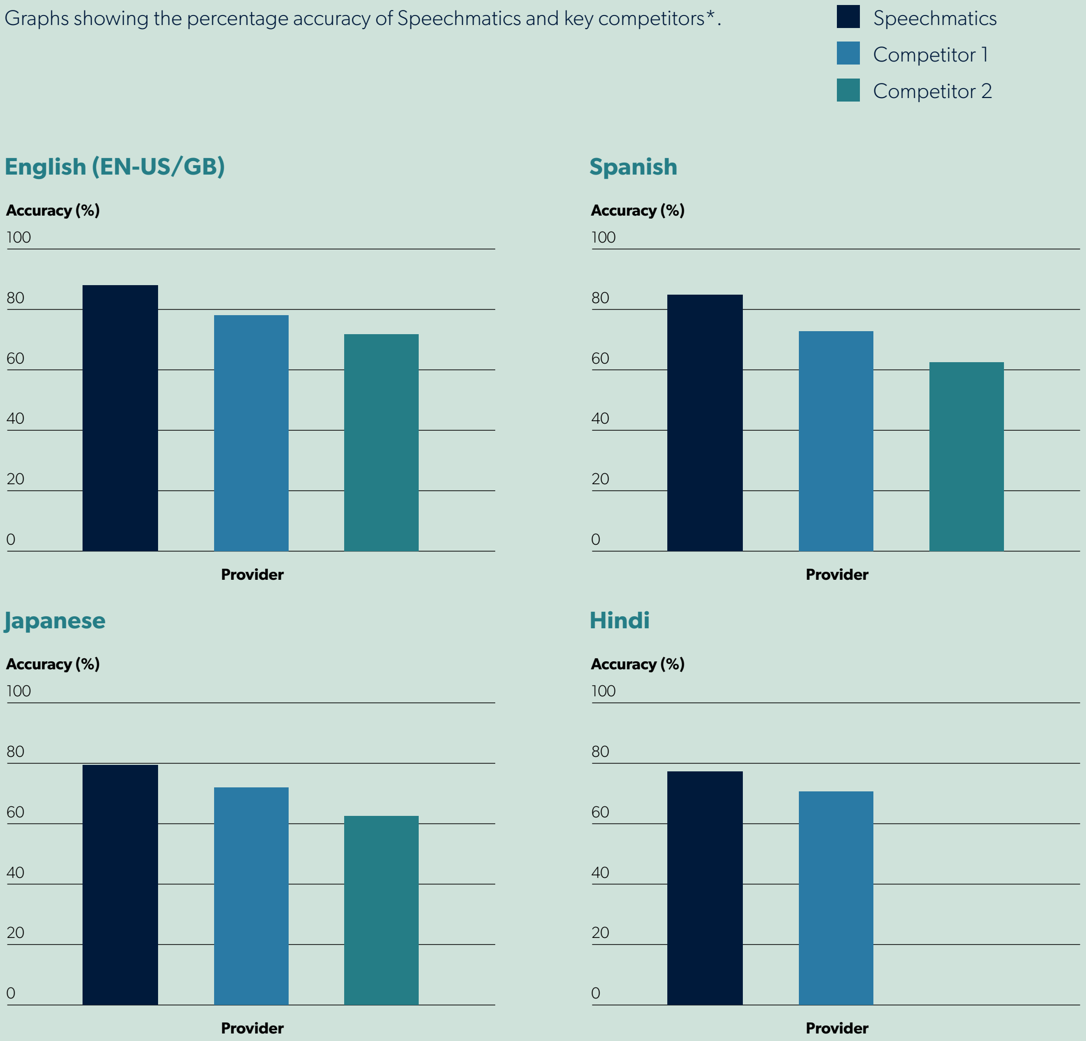
$$\text{Accuracy (\%)} = \frac{100 * (N - E)}{N}$$

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=1. The accuracy is therefore:

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

Figure 3:

Graphs showing the percentage accuracy of Speechmatics and key competitors*.



*Figures taken from 30/08/17–31/08/2017.

Speed of build

AL can build a new language pack in as little as a week, whether in a language we have built before or in a language that is brand new to us.

This means we can respond to the needs of our users quickly and efficiently, by building them the languages they need when they need them, by customising them to use-cases and even by keeping track of languages as they evolve and novel words spread through the world. Our use of machine learning and neural networks makes this possible.

New languages

We aim to make AL as language independent as possible, enabling it to generalise 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.

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 and/or customers 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 13 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.

So… how long does a build take?

Because of the factors outlined previously, we cannot give a general figure for how long each build will take.

However, Hindi and Biology Lecture English in the case studies below are two extremes of build times. Most language pack builds are between these two examples.

1 WEEK ·· **8 WEEKS**

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:

Speed of building Biology Lecture English

A customer required a language pack customised on their specific domain (in English) and provided us with a large quantity of data for the purpose.

1. The data took 10 days to acquire due to (a) the large quantity (b) the provider's slow connection
2. The large volume of data required a lot of computer power
3. Multiple builds for the same customer created many simultaneous demands on hardware
4. English builds do not require specific attention

Factors 1–3 above slowed this build down – but it was still completed in **eight weeks**.

CUSTOMISATION

We have spent a lot of time and effort making sure our models work as generally as possible.

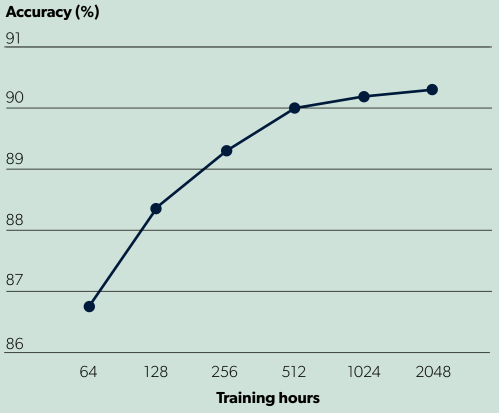
We use cutting-edge techniques to make our speech recognition speaker, accent and recording environment independent. Because of this we are confident our standard language packs provide an out-of-the-box solution with performance matching the needs of almost all use-cases.

However, we do acknowledge that there are some use-cases where very high accuracy of speech recognition is required and the data is markedly different from what we might usually expect. It may have very specialised vocabulary, for example, or an idiosyncratic acoustic environment. In these cases AL offers a means of producing customised language packs.

This is possible because AL can build language packs quickly and requires little intervention to make user data suitable to use.

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.

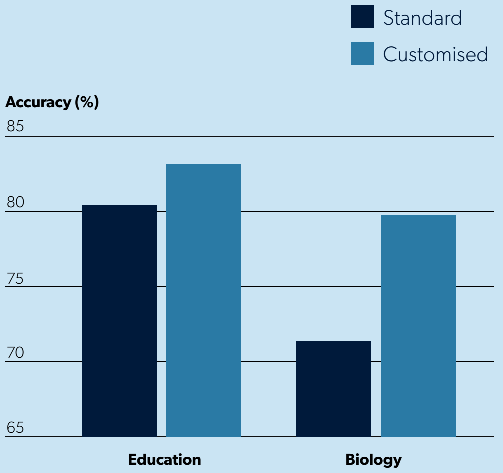


Case study:

Customisation of English for lecture recordings

A customer had a use case for transcribing lectures on a variety of subjects. Due to the unusual vocabulary used in the domains being lectured upon and the distinct acoustic environment of the lecture halls we built models with AL incorporating data from their recordings.

We made builds in two domains – Biology and Education, giving 29% and 14% reductions in transcription errors, respectively, compared to our standard language packs.



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, because 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.

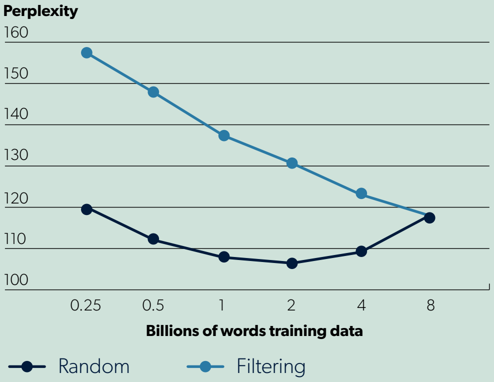
2. Quality

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.

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.



However, high-quality data is not always available, and we must often make do with what is available. This is where, in the traditional world of language pack building, 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.

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 above, 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 are continuously refining our data parameters to match as well as possible how we know our users will use our language packs.

THE CHALLENGES AHEAD

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

All the languages of the world

Our eventual aim is to have a language pack for all the world's languages (and perhaps some of those from out of this world – the sky's the limit!). This is an ambitious aim – it is estimated there are around 7,000 living languages at present, and we hope to cover them all one day.

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. We can then keep confidently rolling out new languages and respond quickly to any user needs without having to invest research effort into how to tackle any particular problem.

And user needs are always increasing. With globalisation 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, especially in regions of the world with lower literacy rates. 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 recognise its power.

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. Our main technologies to date have been privately deployed, including on-premises and batch speech recognition through our cloud transcription service www.speechmatics.com, but in some use-cases this is not appropriate. Real time or embedded systems are required and the language pack requirements in these cases are different.

The Speechmatics Difference

The field of speech recognition is rapidly evolving, with increasing expectations of 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 expertise in machine learning and neural networks to rapidly assimilate the best new methods and advances in the field, AL provides a flexible, evolving framework to provide high quality, fast automatic speech recognition.

SPEECHMATICS

Brookmount Court
Kirkwood Road, Cambridge
CB4 2QH United Kingdom
www.speechmatics.com